Momentum versus Independence in Sequences of Binary Data: Theory and Application to Basketball

Bachelorarbeit
zur Erlangung des akademischen Grades „Bachelor of Science “(B.Sc.)
eingereicht
beim Prüfungsausschuss für den Bachelor-Studiengang
Economics (Politische Ökonomik)
der
Fakultät für Wirtschafts- und Sozialwissenschaften der
Ruprechts-Karls-Universität Heidelberg

2019

Nikolaus Schäfer
geboren in Karlsruhe

Declaration of Authorship

„Hiermit versichere ich, dass ich die vorliegende Arbeit selbständig und ohne unerlaubte fremde Hilfe verfasst habe und dass alle wörtlich oder sinngemäß aus Veröffentlichungen entnommenen Stellen dieser Arbeit unter Quellenangaben einzeln kenntlich gemacht habe.“

Heidelberg, July 3, 2019

Nikolaus Schäfer

Abstract

The hot hand, i.e. the phenomenon that agents who recently experienced a streak of successful outcomes will have a greater chance of successful outcomes in the succeeding attempts (than would be expected by chance), has long been regarded a cognitive illusion. Recent discoveries, primarily by economists Adam Sanjurjo and Joshua B. Miller, have called this claim into question. This thesis provides a summary of the selection bias in the seminal work on the hot hand fallacy by Gilovich et al. (1985), which was discovered by Miller and Sanjurjo (2018b). We will see, upon correcting the bias the conclusion of the canonical paper reverses towards the existence of the hot hand. Furthermore, a numerically tractable calculation (in R) of the size of the bias is provided and discussed extensively. Finally, an approach for testing for the hot hand on a team level (rather than an individual level) is explored. Using the data available to me, I find a small yet significant effect of hot hand shooting on NBA teams during the 2017-2018 NBA season.

Contents

1 Introduction to the Hot Hand

On January 23rd, 2015 the Golden State Warriors played the Sacramento Kings at Oracle Arena in downtown Oakland for an NBA regular season game. During the third quarter the unthinkable happened. Klay Thompson, shooting guard of the Warriors scored 37 points in one quarter while shooting a perfect 13 of 13 from the field (9 of 9 from three). This was (and still is) an NBA record for points in a quarter, 3-pointers in a quarter, and tied a league record for field goals made in a quarter. To put this into perspective, during the 2014-15 NBA regular season Thompson averaged 21.7 points per game (5.4 per quarter) while shooting 46.3 % from the field and 43.9 % from three (see NBA Media Ventures, LLC (2019)). This means Thompson's 37-point quarter exceeded his average by a whopping 585 %.

This performance rekindled an ongoing debate about the existence of the hot hand. As Miller and Sanjurjo (2016) point out, the use of the term "hot hand" shooting is rather vague and complex and varies from paper to paper. The only common denominator is that it refers to a temporary elevation of a player's ability, i.e. the probability of a successful shot. To believe that such a temporary elevation exists from time to time becomes especially reasonable when one considers Thompson's shot sequence from above as a sequence of independent Bernoulli-distributed random variables with $X_i = 1$ if the shot was made and $X_i = 0$ if the shot was missed. Then the chances of hitting 13 shots in a row would be less than 1 in 22000 [1]. However, in 1985, a now very-well known paper was published by the psychologists Thomas Gilovich, Robert Vallone, and Amos Tversky (Gilovich et al. (1985, henceworth GVT)) about the (non-)existence of a "hot hand"-effect [2]. They seemed to provide conclusive evidence that the belief in a hot hand, contrary to widespread conception, is a fallacy. Many studies and papers have built on their work. The *hot hand fallacy* has been used to explain various anomalies and puzzles outside the sports world (see Miller and Sanjurjo (2018b, Footnote 2)). Up until now, many studies have replicated GVT's work in various domains and have found similar results. This lead to an almost three decade long scientific consensus that the existence of the hot hand is a widespread cognitive illusion and no unequivocal evidence for its existence can be found. However, in 2015, the economists Joshua B. Miller and Adam Sanjurjo published the first of many papers in which they explain and mathematically prove that the estimator GVT used throughout their study was actually underlying a downward bias (see Miller and Sanjurjo (2014, 2016, 2018b,a)). Upon correcting the bias they find that the results of GVT actually reverse and provide a strong implication towards the existence of the hot hand.

Section 2 will provide a summary of the findings of Miller and Sanjurjo (2018b). It will discuss the estimator GVT used in their studies and explain the issues Miller and Sanjurjo uncovered in their work. In the first part of Section 3 a detailed version of the algorithm that corrects the bias of GVT's estimator is supplied and an implementation in R is discussed.

[1] $0.463^{13} \approx 4.49 \cdot 10^{-5}$

[2] In particular, their studies examined if there is any kind of momentum/streak shooting in basketball.

In the second part the algorithm is used to correct for the bias in GVT's analysis. Upon adjusting we will see that the conclusion actually reverses, i.e. we find significant evidence for the existence of the hot hand. Section 4 introduces an approach to examine the hot hand effect on a whole team rather than a specific player.

2 A Discussion of the Bias

To understand the findings of Miller and Sanjurjo (2018b) and what it eventually means for the hot hand fallacy, it is first necessary to recapitulate how GVT concluded that the hot hand is a "powerful and widely shared cognitive illusion" (p. 313). This is done in Section 2.1. Section 2.2 explains the nature of the bias on an intuitive level, while Section 2.3 then dives into the more abstract and general notation.

2.1 GVT's Approach

As pointed out by Miller and Sanjurjo (2016) a general problem that arises when trying to study the hot hand is that it is impossible for the researcher to know for sure when a player is actually in the hot state. Hence, an operational definition is required. GVT designed a test (which they refer to as *Analysis of Conditional Probabilities*) for the following definition of the hot hand[3]:

Definition 2.1. `Hot Hand (GVT)`
A player has a **hot hand/is under the influence of the hot-hand effect**, if his/her probability of hitting a shot is greater after a streak of successful shots than after a streak of misses. In other words if

$$P(\text{hit} \mid \text{streak of hits}) - P(\text{hit} \mid \text{streak of misses}) > 0.$$

Hereby a streak is defined as:

Definition 2.2. `Streak`
Let S be the set of shot indices and $\{x_s\}_{s\in S}$ the sequence of shot outcomes, where $x_s = 1$ if shot s is a hit, and $x_s = 0$ if it is a miss. A **streak** occurs at shot s if a player has just completed k ($k \in \mathbb{N}$) or more shots with the same outcome, i.e. if $x_{s-1} = x_{s-2} = \ldots = x_{s-k}$.

They applied a test, using this definition, to shot sequences in three different contexts. I will now provide a short summary of the three different studies and how the test was applied in each scenario. In the first study, NBA field goal data were considered. They were obtained

[3] GVT actually considered two more definitions/conducted two other tests, a Wald-Wolfowitz run test and a test of fit which they refer to as *Test of Stationarity*. However, Wardrop (1999) showed that the *Analysis of Runs* and the *Analysis of Conditional Probabilities* basically amount to the same test (correlation coefficient of -0.993). Moreover he exposed some serious flaws in the test of stationarity and concluded that it „should not be used "(p.4). Therefore, this thesis will only focus on GVT's *Analysis of Conditional Probabilities.*

from the statistician of the Philadelphia 76ers, who kept field goal records of individual players of 48 home games during the 1980-81 season. GVT examined if players hit a higher percentage of their shots after having just made their last shot (or last several shots) compared to having missed their last shot (or last several shots)[4]. The second study was about NBA free throw data from the Boston Celtics during the 1980-81 and 1981-82 season. Specifically, all trips to the line where exactly two free throws were shot were considered [5]. They explored whether a player is more likely to hit their second free throw after making the first compared to missing the first. Thirdly in a controlled shooting experiment, GVT asked 26 NCAA Division I[6] players (12 men and 14 female) to shoot 100 shots from a distance where they normally make around 50 % of their shots. Again, they compared the probability of a hit conditioned on making the previous k shots to the probability of a hit conditioned on missing the previous k shots ($k = 1, 2, 3$). The third study was made in order to eliminate in-game factors such as defensive pressure, shot distance and fatigue which undoubtedly play a factor in studies one and two. It is clear to see that GVT used the estimator $\hat{D}_k^i = \hat{P}^i(\text{hit} \mid k \text{ hits}) - \hat{P}^i(\text{hit} \mid k \text{ misses})$ in all three contexts. $\hat{P}^i(\text{hit} \mid k \text{ hits/misses})$ denotes the probability of a successful shot of player i conditional on the fact that player i hit/missed the k immediately preceding shots [7]. They then argued that the higher $\hat{D}_k^i$ the stronger the evidence that player i occasionally has a “hot hand”. As a result they performed a two sample t-test for the shot sequences in each of the above described scenarios using the null hypothesis $\mathbb{E}[\hat{P}^i(\text{hit} \mid k \text{ hits}) - \hat{P}^i(\text{hit} \mid k \text{ misses})] = 0$. To the untrained eye, this approach might seem reasonable and intuitively correct. However, Miller and Sanjurjo (2018b) show that against any intuition $\mathbb{E}[\hat{P}^i(\text{hit} \mid k \text{ hits}) - \hat{P}^i(\text{hit} \mid k \text{ misses})]$ is actually substantially **smaller than zero** for sequence lengths used in empirical studies (see Figure 2 in Section 2.3). Therefore, it is no wonder that GVT found $\hat{P}^i(\text{hit} \mid k \text{ hits}) - \hat{P}^i(\text{hit} \mid k \text{ misses})$ to be statistically significant greater than zero for only one of 26 players in study three. Since this downward bias in GVT’s approach was undetected for nearly three decades, it goes without saying that it is not easy to identify where exactly GVT’s (and maybe oneself’s) thinking went wrong. The following subsection will try to provide an intuitive approach to understanding this fallacy [8].

2.2 Where Does the Bias Come From?

To adequately explain the nature of the bias in GVT’s approach, this subsection gives an example where the origin of the bias is (more or less) directly visible. The next subsection then presents a more abstract notation where it becomes evident that the bias exists in all scenarios

[4] GVT examined hit/miss streaks from lengths one to three or to use Definition 2.2 they considered $k = 1, 2, 3$.

[5] shooting exactly two free throws is the most common case in basketball. However, shooting one free throw or even three free throws during a single trip is also possible.

[6] The NCAA Division I is the highest level of American college sports where student athletes compete in various kinds of sports for their respective colleges.

[7] In the case of study 1, the preceding shots had to occur in the same game.

[8] Ironically, as Miller and Sanjurjo (2018b) wisely point out, upon their findings, the hot hand fallacy itself can be viewed as a fallacy.

no matter what the underlying baseline probability $p \in (0,1)$ of a successful shot, or the streak length $k \in \mathbb{N}$ might be.

Example 2.3. (Coin flipping)
This example is based on the coin flip example, given in Miller and Sanjurjo (2018b) or Miller and Sanjurjo (2016). This thesis will shine a little more light on the exact math behind this example as there are many who doubted the correctness of the math at first (see Fisher, Sam (2015) or Gelman, Andrew (2016)).
Consider flipping a fair coin 100 times and writing down the outcome of each flip on a sheet of paper. Upon completing the hundred flips you underline all the flips that were preceded by a heads flip. Of course, you now expect the proportion of head flips in all the underlined flips to be one-half. Counter-intuitively, this is incorrect. The expected proportion of heads conditional on the fact that the immediately preceding flip was a heads is actually 49.49% in this case [9]. Why is that so? To understand why the expected proportion of heads in all the underlined flips is strictly less than 0.5, it might be easier to consider a simpler case where you only flip the coin four times. That means there are $2^4 = 16$ possible head-tail sequences. Since the flips are independent, each sequence is equally likely to occur with a probability of $0.5^4 = \frac{1}{16} = 0.0625$. Column 1 of Table 1 lists all of those sequences.

To calculate the proportion of heads on flips that were immediately preceded by a head flip, one may just calculate the following conditional expectation:

$$\mathbb{E}[\text{prop. of Hs on flips preceded by an H} \mid \text{a flip was actually underlined}]$$

Notice that it only makes sense to observe the proportion of heads if there was actually a flip recorded (underlined). That is of course the case, when there exists a head flip in our sequence that was not the last flip (meaning the flips TTTT and TTTH do not meet this criterion).
Calculating this expectation is now straightforward. Define the random variable $Y :=$ proportion of heads on flips preceded by an H, and the event $\mathcal{H} :=$ The sequence contains a head flip that is not the last flip, i.e. $\{\text{TTTT, TTTH}\}^C$. Then:

$$\mathbb{E}[Y|\mathcal{H}] = \frac{\sum\limits_{\omega \in \mathcal{H}} Y(\omega)}{|\mathcal{H}|} = \frac{17}{3} \cdot \frac{1}{14} = \frac{17}{42} < \frac{1}{2},$$

where $Y(\omega)$ can be taken from Column 4 of Table 1 and $|\mathcal{H}| = 16 - |\{\text{TTTT, TTTH}\}| = 14$. The important thing to notice here is that the sequence (rather than the flip) is the primitive outcome in this case. This is the reason why the weight the (conditional) expectation places on each sequence's associated proportion is independent of the number of recorded flips.

[9] see Section 3 for more details on this calculation

Table 1: `The Bias in the Case of Four Coin Flips`

Sequence	**# of Hs after an H flip**	**# Ts after an H flip**	**Proportion of Hs = $X(\omega)$**
TTTT	-	-	-
TTTH	-	-	-
TTH$\underline{\text{T}}$	0	1	0
TH$\underline{\text{T}}$T	0	1	0
H$\underline{\text{T}}$TT	0	1	0
TTH$\underline{\text{H}}$	1	0	1
TH$\underline{\text{T}}$H	0	1	0
H$\underline{\text{T}}$H$\underline{\text{T}}$	0	2	0
H$\underline{\text{HT}}$T	1	1	$\frac{1}{2}$
TH$\underline{\text{HT}}$	1	1	$\frac{1}{2}$
H$\underline{\text{T}}$TH	0	1	0
TH$\underline{\text{HH}}$	2	0	1
H$\underline{\text{T}}$H$\underline{\text{H}}$	1	1	$\frac{1}{2}$
H$\underline{\text{HT}}$H	1	1	$\frac{1}{2}$
H$\underline{\text{HHT}}$	2	1	$\frac{2}{3}$
H$\underline{\text{HHH}}$	3	0	1
Sum	**12**	**12**	$\boldsymbol{\frac{17}{3}}$

Notes: Column 1 list all the possible sequence one can obtain by flipping a coin four times. For each sequence that has an H during the first three flips, Columns 2 and 3 count the number of underlined Hs and Ts respectively. Column 4 then represents the proportion of heads, which is just the quotient of Column 2 and the sum of Columns 2 and 3.

For more intuition about the bias and its connection to other economic paradoxes see also Miller and Sanjurjo (2016, 2017). A more hands-on approach is to look at simulations of coin flips for oneself. In Schäfer (2019) I provide an R program which simulates coin flips with variable trial length n, streak length k, and success probability p. I invite the reader to try it out.

2.3 The Selection Bias in the General Case

After seeing some intuition for the bias in an experiential example, this subsection is dedicated to discussing the bias in the general case. Consider $\mathbf{X} = (X_i)_{i=1,\ldots,n}$, a sequence of Bernoulli-distributed random variables. I call $X_i = 1$ a "success" (i.e. a successful shot) and $X_i = 0$ a "failure" (i.e. a missed shot). Let the probability of a success be denoted by p. Now, considering such a sequence of length n, we are interested in the expected proportion of successes following k successive successes. Therefore, to calculate the proportion, it is necessary to first filter out all the trials that follow k successive successes.

Definition 2.4.

Let $\mathbf{X} = (X_i)_{i=1,\dots,n}$ be a sequence of binary random variables, with $X_i = 1$ a "success" and $X_i = 0$ a "failure". Then we can define the subset of trials that immediately follow k consecutive successes as

$$I_k(\mathbf{X}) := \left\{ i : \prod_{j=i-k}^{i-1} X_j = 1 \right\} \subseteq \{k+1, \dots, n\}$$

The following example helps to illustrate this abstract notation.

Example 2.5.

Consider the sequence $\mathbf{X} = 101111$ of length 6. Then:

$$\begin{aligned} I_1(\mathbf{X}) &= \left\{ i : \prod_{j=i-1}^{i-1} X_j = 1 \right\} = \{2, 4, 5, 6\} \subseteq \{2, \dots, 6\} \\ I_2(\mathbf{X}) &= \left\{ i : \prod_{j=i-2}^{i-1} X_j = 1 \right\} = \{5, 6\} \subseteq \{3, \dots, n\} \\ I_3(\mathbf{X}) &= \left\{ i : \prod_{j=i-3}^{i-1} X_j = 1 \right\} = \{6\} \subseteq \{4, \dots, n\} \end{aligned}$$

After computing the subset of trials that follow k consecutive successes, the proportion of successes is just the the number of successes ($\sum_{i \in I_k(\mathbf{X})} X_i$) in this subset divided by the number of trials, i.e. the cardinality of $I_k(\mathbf{X})$. Theorem 2 of Miller and Sanjurjo (2018b) now shows that the expected proportion of successes in $I_k(\mathbf{X})$ is actually less than the success probability p. Theorem 2.6 replicates this theorem.

Theorem 2.6.

Let $\boldsymbol{X} = (X_i)_{i=1,\dots,n}$, $n \geq 3$, be a sequence of independent Bernoulli trials, each with probability of success $0 < p < 1$. Let $\hat{P}_k(\boldsymbol{X})$ be the proportion of successes, that is,

$$\hat{P}_k(\boldsymbol{X}) := \frac{\sum\limits_{i \in I_k(\boldsymbol{X})} X_i}{|I_k(\boldsymbol{X})|}.$$

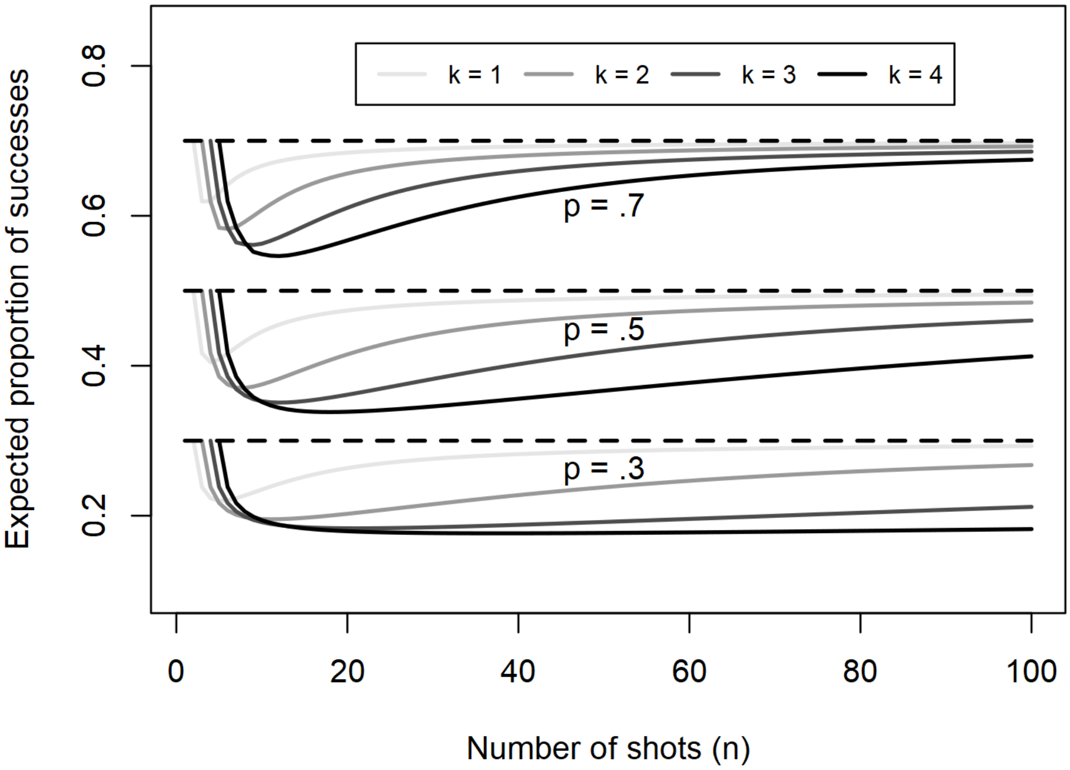
$\hat{P}_k$ is a biased estimator of $\mathbb{P}\left(X_t = 1 \mid \prod\limits_{j=t-k}^{t-1} X_j = 1 \right) \equiv p$ [10] for all k such that $1 \leq k \leq n-2$. In particular,

$$\mathbb{E}[\hat{P}_k(\boldsymbol{X}) | I_k(\boldsymbol{X}) \neq \emptyset] < p. \tag{1}$$

[10] Note that since all the trials are independent it does not matter what outcomes the k previous trials had. Therefore $\mathbb{P}\left(X_t = 1 \mid \prod_{j=t-k}^{t-1} X_j = 1 \right) = \mathbb{P}(X_t = 1) = p$

For a proof see Appendix A.1 of Miller and Sanjurjo (2018b). Notice that $I_k(\mathbf{X}) \neq \emptyset$ is just the condition that a flip was actually underlined in Example 2.3 ($I_1(\text{TTTT}) = I_1(\text{TTTH}) = \emptyset$). Although $\hat{P}_k(\mathbf{X})$ is biased, Miller and Sanjurjo (2018b) show that it is a consistent estimator, meaning $\hat{P}_k(\mathbf{X})$ converges in probability to p. Nevertheless, Figure 1 shows for different streak lengths and success probabilities that the bias stays quite substantial even for more than 100 trials (the trial length GVT chose for their third study). Generally speaking, one can observe that the bias is greater for longer streak lengths. This observation will become very important in Section 3.

Figure 1: Bias in the Expected Proportion of Successes



Notes: $\mathbb{E}[\hat{P}_k(\mathbf{X}) \mid k \text{ successes}]$ as a function of the total number of trials n for different streak lengths ($k = 1, \ldots, 4$) and different probabilities of success ($p = 0.3, 0.5, 0.7$). The used algorithm is provided in Appendix R (Algorithm 4). For a discussion of the algorithm, see Section 3.1.

What's left now is to connect these findings with GVT's approach discussed in 2.1 and their assumption that $\hat{D}^i_k := \hat{P}^i(\text{hit}|k \text{ hits}) - \hat{P}^i(\text{hit}|k \text{ misses})$ is expected to be zero [11].
It has become evident that for a player i $\hat{P}^i(\text{hit}|k \text{ hits})$ is actually expected to be smaller than p^i. By symmetry it follows that $P^i(\text{hit}|k \text{ misses})$ is expected to be greater than p^i [12]. To use

[11] This thesis follows the notation suggested by Miller and Sanjurjo (2018b) and uses $\hat{P}^i(\text{hit}|k \text{ hits})$ for both the random variable $\hat{P}_k(\mathbf{X})$ and its realization $\hat{P}_k(\mathbf{x})$ in order to facilitate the comparison with GVT's analysis. In similar fashion, $\hat{P}^i(\text{hit}|k \text{ misses})$ is used for the proportion of successes on trials that immediately follow k consecutive failures.

[12] For a visualization of the symmetry, see Figure 5 in Appendix G

the notation of Theorem 2.6, this means that $\mathbb{E}[1 - \hat{Q}_k(\mathbf{X})|J_k(\mathbf{X}) \neq \emptyset] > p$, where $\hat{Q}_k(\mathbf{X})$ is the proportion of failures on the subset of trials that immediately follow k consecutive failures, $J_k(\mathbf{X}) := \Big\{j : \prod\limits_{i=j-k}^{j-1}(1 - X_i) = 1\Big\} \subseteq \{k+1, \ldots, n\}$.
For the difference in the probability of success when comparing trials that immediately follow k consecutive successes with trials that immediately follow k consecutive failures ($D_k := \mathbb{P}(X_t = 1 \mid \prod\limits_{j=t-k}^{t-1} X_j = 1) - \mathbb{P}(X_t = 1 \mid \prod_{j=t-k}^{t-1}(1 - X_j) = 1)$), GVT used the estimator

$$\hat{D}_k(\mathbf{X}) := \hat{P}_k(\mathbf{X}) - [1 - \hat{Q}(\mathbf{X})]. \tag{2}$$

It should be clear by now that $\hat{D}_k$ is a biased estimator of D_k. Theorem 2.7 replicates Theorem 3 of Miller and Sanjurjo (2018b) and shows the existence of a downward bias.

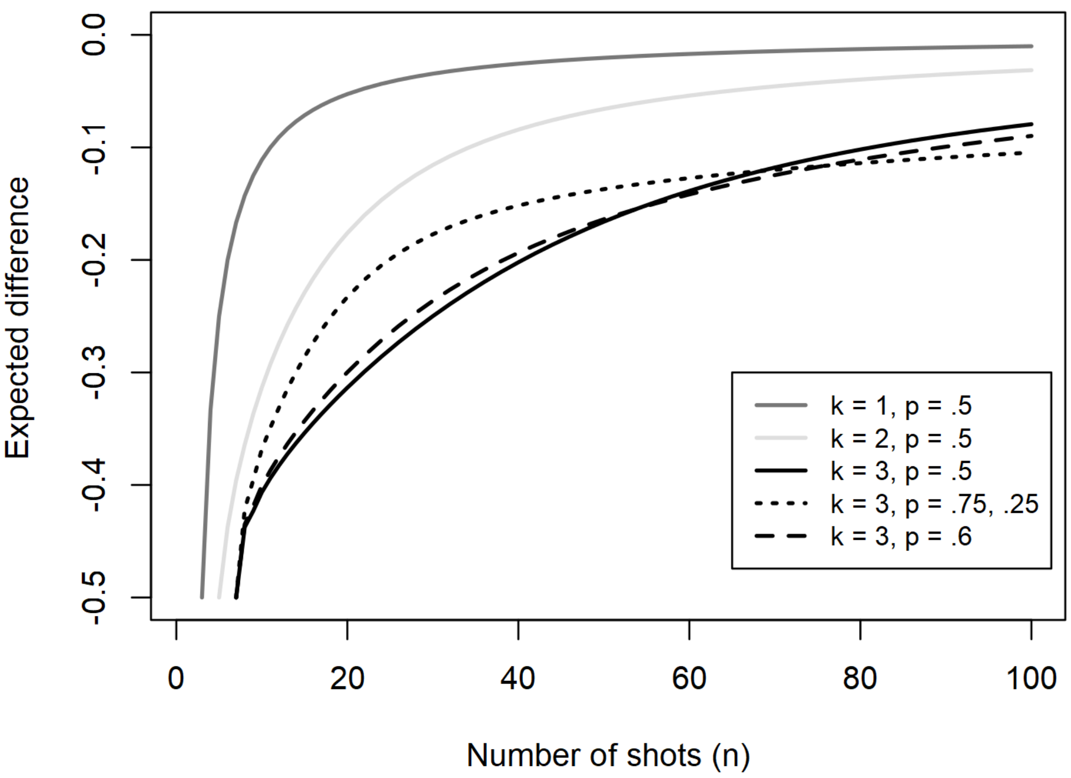
Theorem 2.7.
Let $\boldsymbol{X} = (X_i)_{i=1,\ldots,n}$, $n \geq 3$, be a sequence of independent Bernoulli trials, each with probability of success $p \in (0,1)$. Let $\hat{P}_k(\boldsymbol{X})$ be the proportion of successes on the subset of trials $I_k(\boldsymbol{X})$ that immediately follow k consecutive successes, and $\hat{Q}(\boldsymbol{X})$ be the proportion of failures on the subset of trials $J_k(\boldsymbol{X})$ that immediately follow k consecutive failures. $\hat{D}_k(\boldsymbol{X}) := \hat{P}_k(\boldsymbol{X}) - [1 - \hat{Q}_k(\boldsymbol{X})]$ is a biased estimator of $D_k := \mathbb{P}(X_t = 1 \mid \prod_{j=t-k}^{t-1} X_j = 1) - \mathbb{P}(X_t = 1 \mid \prod_{j=t-k}^{t-1}(1 - X_j) = 1 \equiv 0$ for all such k such that $1 \leq k < n/2$. In particular,

$$\mathbb{E}[\hat{D}_k(\boldsymbol{X}) \mid I_k(\boldsymbol{X}) \neq \emptyset, J_k(\boldsymbol{X}) \neq \emptyset] < 0 \tag{3}$$

For a proof see Appendix A.4.1 of Miller and Sanjurjo (2018b). Notice that the proof of this theorem does not follow directly out of 2.6. For a sequence $\mathbf{x} \in \{0,1\}^n$ the difference in proportions is well defined if, and only if $I_k(\mathbf{x}) \neq \emptyset$ and $J_k(\mathbf{x}) \neq \emptyset$. The well-definiteness of the respective proportions, however, is looser. For the proportion of successes that follow k successes only $I_k(\mathbf{x}) \neq \emptyset$ has to be satisfied, and for the proportion that follows k failures only $J_k(\mathbf{x}) \neq \emptyset$. Therefore, the set of sequences for which the difference in proportions is well-defined is a strict subset of the sequences for which either of the proportions itself are well-defined. Nevertheless, the argument of the proof remains similar.

Figure 2 shows the bias of $\hat{D}_k^i$ for several streak lengths k and success probabilities p. Again, we see that the bias remains quite substantial for a trial length of greater than 100. For the same p and n the bias increases with increasing streak length k. For specific cases, the bias is even greater than 10 percentages points.

Figure 2: Bias in the Expected Difference of Proportions



Notes: $\mathbb{E}[\hat{D}_k \mid k$ successes] as a function of the total number of trials n for different streak lengths ($k = 1, 2, 3$) and different probabilities of success p. The used algorithm is provided in Appendix R (Algorithm 6). For a discussion of the algorithm, see Section 3.1.

3 The Size of the Bias and Its Implications for the Hot Hand

In Section 2 I discussed the nature of the bias. In order to see the implications that this bias puts on the interpretation of studies of the hot hand using Definition 2.1, it is necessary to quantify the bias. As it turns out, computing the bias for a sequence of length n, success probability p, and streak length k is far from trivial, especially when n increases. In the first part of this section, I will discuss the algorithm provided by Miller and Sanjurjo (2018a, Appendix E) and my implementation of it in R. In Section 3.2 I will discuss and de-bias the results of GVT using the methods provided by Miller and Sanjurjo (2016, 2018b) and then examine the consequences for the hot hand fallacy.

3.1 Algorithm for Calculating the Size of the Bias

In Appendix E Miller and Sanjurjo (2018a) provide a method to calculate the exact sampling distributions of both the proportion and the difference in proportions. Those sampling distributions can then be used to calculate the exact values of $\mathbb{E}[\hat{P}_k(\mathbf{X})|I_k(\mathbf{X}) \neq \emptyset]$, $\mathbb{E}[1 - \hat{Q}_k(\mathbf{X})|J_k(\mathbf{X}) \neq \emptyset]$, and $\mathbb{E}[\hat{D}_k \mid I_k\mathbf{X}) \neq \emptyset, J_k(\mathbf{X}) \neq \emptyset]$. This thesis will explain the procedure of the computation in more detail and provide an actual executable numerically tractable code (in R) for all of the above. This extends the work of Miller and Sanjurjo (2018a) as they only provide a pseudo-code for the sampling distribution of $\mathbb{E}[\hat{P}_k(\mathbf{X})|I_k(\mathbf{X}) \neq \emptyset]$ and no code for the used helper functions, $\mathbb{E}[1-\hat{Q}_k(\mathbf{X})|J_k(\mathbf{X}) \neq \emptyset]$, and $\mathbb{E}[\hat{D}_k \mid I_k\mathbf{X}) \neq \emptyset, J_k(\mathbf{X}) \neq \emptyset]$ at all. The coded functions are used to create the plots in Figures 1-3, Figures 5-7, and also later to de-bias the GVT's estimator. The whole code can be found in Schäfer (2019). A summary of the algorithms is also provided in Appendix R.

3.1.1 Expected Proportion

Recall Definition 2.4 and Equation (1) from Theorem 2.6. For a given trial number n, streak length k, and success probability p we are interested in calculating the expected proportion of successes on trials that immediately follow k consecutive successes/failures. In the following I will focus on the proportion of successes on trials that immediately follow k consecutive successes as the other case (trials following k consecutive failures) works analogously.

For an arbitrary binary sequence $\mathbf{x} \in \{0,1\}^n$ I use the notation of Miller and Sanjurjo (2018a) and rewrite $\hat{P}_k(\mathbf{x})$ as

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

where $M_k^0(\mathbf{x}) = \sum_{i \in I_k(\mathbf{x})}(1-x_i)$ and $M_k^1(\mathbf{x}) = \sum_{i \in I_k(\mathbf{x})} x_i$ are the number of failures/successes on trials following k consecutive successes. In the following the k is suppressed to ease notation. $\mathbb{E}[\hat{P}_k(\mathbf{x}) \mid I_k(\mathbf{x})]$ is now uniquely determined by the joint probability distribution of counts $\mathbb{P}((M^0(\mathbf{X}), M^1(\mathbf{X})) = (m^0, m^1))$. Intuitively speaking, the idea is to calculate the probability for any $M_k^0(\mathbf{x})$, $M_k^1(\mathbf{x})$ combination possible in n trials and streak length k. In the following I will call a unique $M_k^0(\mathbf{x})$, $M_k^1(\mathbf{x})$ combination a *count realization.*

Finding a function, whose output is the exact joint distribution, is quite challenging in the general case. Miller and Sanjurjo (2018a) provide a recursively defined algorithm which uses *dictionaries.* A *dictionary* is a data structure, which is used to store a group of objects. It is characterized by a set of *keys* and a set of *values*, such that each key has a single associated value and each element of a dictionary can be written as a pair: (key: value). Connecting this principle to our case at hand, it comes natural that the set of keys consists of the unique count realizations and the set of values of the associated probabilities. The following Definition summarizes the just made thought progress and introduces some notation.

Definition 3.1.
For a sequence $\mathbf{x} \in \{0,1\}^n$ and streak length k, the joint distribution,

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

can be represented by a dictionary

$$D := (\mathbf{m} : p_D(\mathbf{m}))_{\mathbf{m} \in D_c},$$

where $\mathbf{m} := (m^0, m^1)$ is a unique count realization and $D_c := \{\mathbf{m} \in \mathbb{N}^2 \mid p_D(\mathbf{m}) > 0\}$ is the set of count realizations with non-zero probability. A unique pair, $(\mathbf{m} : p_D(\mathbf{m}))$, is called a *count-probability pair.*

Before I move on to the general computation and the algorithm, let us first look at an example to get a sense of intuition about the procedure.

Example 3.2.
Let us consider a slightly more generalized version of Example 2.3, i.e. $n = 3$, $k = 1$ but $p \in (0, 1)$ and not fixed. Define $q := 1 - p$, the probability of a failure. Table 2 replicates Table E.I of Miller and Sanjurjo (2018a) and reports the sampling distribution as well as the corresponding dictionary for the above mentioned case. Notice that since p is not fixed to 0.5, I have replaced the H and T notation from Example 2.3 with the more standard 1 and 0 notation for a success and failure. In the table to the left Column 1 lists all the possible sequences of length 3, where all trials that follow a success are underlined just like in Table 1. Column 2 reports the respective occurrence probabilities. Column 3 lists the count realizations. Therefore, the left entry corresponds to the number of underlined failures (zeroes) and the right entry to the number of underlined successes (ones). Now we can form our dictionary (the right table). Column 1 of the right table lists all the unique count realizations found in Column 3 of the left table. Each unique count $\mathbf{m} = (m^0, m^1)$ has a uniquely associated probability, which is just the sum of the probabilities of all the sequences with the same associated count. So for $\mathbf{m} = (0, 0)$, $p_D(\mathbf{m}) = q^3 + q^2p = q^3 + q^2(1 - q) = q^3 + q^2 - q^3 = q^2$ (see Rows 1 and 2 of the left table).

From the dictionary one can now easily compute the expected proportion in the conventional manner:

$$\mathbb{E}[\hat{P}_k(\mathbf{x}) \mid I_k(\mathbf{x}) \neq \emptyset] = \sum_{\mathbf{m} \in D_c^*} \frac{m^1}{m^0 + m^1} p_D^*(\mathbf{m}), \tag{4}$$

where $D_c^* = D_c \backslash \{(0, 0)\}$ and $p_D^*(\mathbf{m}) := p_D(\mathbf{m}) / \sum_{\mathbf{m}' \in D_c^*} p_D(\mathbf{m}')$. The count $(0, 0)$ is of course excluded from the computation since there was no trial underlined in the corresponding sequences. Hence, the definition of D_c^* and $p_D^*(\mathbf{m})$.

Now, since

$$(q + q^2)p + qp^2 + qp^2 + p^3 = ((1 - p) + (1 - p)^2)p + 2(1 - p)p^2 + p^3 = 2p - p^2,$$

Table 2: Dictionary Representations of Count-Probability Pairs

Sample space of sequences		
Sequence	**Probability**	**Count Realization**
000	q^3	(0,0)
001	q^2p	(0,0)
$01\underline{0}$	q^2p	$(1,0)$
$1\underline{0}0$	q^2p	$(1,0)$
$01\underline{1}$	qp^2	$(0,1)$
$1\underline{0}1$	qp^2	$(1,0)$
$1\underline{10}$	qp^2	$(1,1)$
$1\underline{11}$	p^3	$(0,2)$

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

Notes: In the table to the left Column 1 displays all the possible outcomes of a binary sequence of length 3. All the trials after a success are underlined. Column 2 reports the corresponding occurrence probabilities, while p is the success and q the failure probability. Finally, Column 3 of the left table lists the number of (failures, successes) that are underlined in Column 1. The table to the right reports the dictionary. Column 1 represents the list of keys (unique count realizations) and Column 2 the list of values (probabilities of the count realizations).

we get

$$\mathbb{E}[\hat{P}_k(\mathbf{x}) \mid I_k(\mathbf{x}) \neq \emptyset] = \frac{1}{1}\frac{qp^2}{2p-p^2} + \frac{1}{2}\frac{qp^2}{2p-p^2} + \frac{2}{2}\frac{p^3}{2p-p^2} = \frac{3p^2-p^3}{4p-2p^2} = \frac{p(3-p)}{2(2-p)}.$$

Therefore, for the case of $p = 0.5$,

$$\mathbb{E}[\hat{P}_1(\mathbf{x}) \mid I_1(\mathbf{x}) \neq \emptyset] = \frac{0.5 \cdot 2.5}{3} = \frac{5}{12}.$$

In practice, often cases are considered, where n and k are significantly higher. In this context, the computation of the dictionary cannot be done by hand as easily as in the above example. For $n = 10$, for instance, one already has to consider a total of $2^{10} = 1024$ sequences. An algorithm applicable to the general case is needed. The following definition provides the basis for the algorithm discussed in this thesis.

Definition 3.3.
For $n \geq 2$, streak length k, $\ell \leq k$ and $r \leq n$, let $D(\ell, r)$ be the dictionary that includes the count-probability pairs for the remaining r trials of a sequence that has $\ell \leq k$ consecutive successes immediately preceding the current trial.

This means, for a given trial length n the dictionary $D(2,1)$ contains all the count-probability pairs for a sequence that has one remaining trial and two successes immediately preceding the current trial. The sequence would look something like this

$$...011?$$

where the dots stand for an arbitrary binary sequence of length $n-4$, and the question mark stands for the nth element of the sequence. For $k=1$ the dictionary would take the form

$$D(2,1) = ((1,0):q,(0,1):p)$$

as the one remaining trial can either be a success (with probability p) or a failure (with probability q). We will see that the dictionary of interest in the general case is $D(0,n)$, i.e. n trials remaining and no preceding successes [13]. Let us look at a simple example.

Example 3.4. Let n be large enough ($n \geq 3$) and $k=1$. Then:

$$D(0,0) = D(1,0) = ((0,0):1),$$

since when zero trials remain in the sequence the only possible count is $(0,0)$, which then occurs with probability 1. We also have

$$D(0,1) = ((0,0):1),$$

as when only one trial remains and the previous trial was a failure, then the only possible count is $(0,0)$. Lastly, we have

$$\begin{aligned} D(1,1) &= D(2,1) = ((1,0):q,(0,1):p) \\ D(1,2) &= ((1,0):q^2+qp,(1,1):pq,(0,2):p^2) \\ D(0,2) &= ((0,0):q^2+qp,(1,0):pq,(0,1):p^2) \\ D(0,3) &= ((0,0):q^2,(1,0):(q+q^2)p,(0,1):qp^2,(1,1):qp^2,(0,2):p^3) \end{aligned}$$

There are two main takeaways from the example. Firstly, $D(0,3)$ is equal to the dictionary in Table 2 and secondly, $D(\ell,0) = ((0,0):1)$ for $0 \leq \ell \leq k$.
It has become clear that if we want to compute the joint distribution of counts, we have to compute the dictionary $D(0,n)$. The crux is now that the dictionaries $D(\ell,r)$ can be defined recursively for $r>0$ and $0 \leq \ell \leq k$, and take the following form:

$$D(\ell,r) = \begin{cases} D(0,r-1)^{(0,0):q} \uplus D(\ell+1,r-1)^{(0,0):p}, & \text{if } \ell < k, \\ D(0,r-1)^{(1,0):q} \uplus D(k,r-1)^{(0,1):p}, & \text{if } \ell = k, \end{cases} \tag{5}$$

where:

[13] This is obvious, since a preceding success would require us to look at a sequence of length greater than n.

(i) $D^{\mathbf{m}':p'} := (\mathbf{m}+\mathbf{m}' : p_D(\mathbf{m}) \cdot p')_{\mathbf{m}\in D_c}$. This means, to each key $\mathbf{m} \in D_c$ $\mathbf{m}'$ is added and each value (probability) $p_D(\mathbf{m})$ is multiplied by p'.

(ii) given two dictionaries A and B, $A \uplus B := (\mathbf{m} : (p_A + p_B)(\mathbf{m}))_{\mathbf{m}\in A_c \cup B_c}$. This means, the new dictionary $(A \uplus B)$ consists of all keys in A and B. If A and B share a key $(\mathbf{m} \in A \cap B)$, the probabilities are added. If $\mathbf{m} \in A$ but not in B (or vice versa), then the probability just takes on $p_A(\mathbf{m})$ (or $p_B(\mathbf{m})$ if it is the other way around).

Algorithms 1-3 in Appendix R describe the helper function (i), the helper function (ii) and the recursive procedure in (5) respectively. Example 3.5 shows the recursive procedure for the situation in Example 3.2.

Example 3.5.
Let $n = 3$, $k = 1$. Our goal is to compute the dictionary D_c from Example 3.2. We have seen that $D_c = D(0,3)$. Using the recursive procedure (5), we get:

$$D(0,3) = D(0,2)^{(0,0):q} \uplus D(1,2)^{(0,0):p}.$$

Similarly, $D(0,2) = D(0,1)^{(0,0):q} \uplus D(1,1)^{(0,0):p}$ and $D(1,2) = D(0,1)^{(1,0):q} \uplus D(1,1)^{(0,1):p}$. Continuing this process we get:

$$\begin{aligned} D(0,1) &= D(0,0)^{(0,0):q} \uplus D(1,0)^{(0,0):p} \text{ and} \\ D(1,1) &= D(0,0)^{(1,0):q} \uplus D(1,0)^{(0,1):p}. \end{aligned}$$

Remember that $D(\ell,0) = ((0,0) : 1)$ for $0 \leq \ell \leq k$. Using this, we get:

$$\begin{aligned} D(0,1) &= ((0,0) : q) \uplus ((0,0) : p) = ((0,0) : 1) \\ D(1,1) &= ((1,0) : q) \uplus ((0,1) : p) = ((1,0) : q, (0,1) : p) \\ \Rightarrow D(0,2) &= ((0,0) : q), (1,0) : qp, (0,1) : p^2) \quad \text{and} \\ D(1,2) &= ((1,0) : q, (1,1) : qp, (0,2) : p^2) \\ \Rightarrow D(0,3) &= ((0,0) : q^2, (1,0) : q^2p, (0,1) : qp^2) \uplus ((1,0) : qp, (1,1) : qp^2, (0,2) : p^3) \\ &= ((0,0) : q^2, (1,0) : (q+q^2)p, (0,1) : qp^2, (1,1) : qp^2, (0,2) : p^3), \end{aligned}$$

which is just what we wanted.

What is left is the calculation of the expected proportion. This can be done in the conventional manner and even in the general case works analogously to Equation (4).
It has been established that $D_c = D(0,n)$. Define $D_c^* := D_c \backslash \{(0,0)\}$, the subset of count

realizations for which the corresponding sequences fulfill $I_k(\mathbf{x}) \neq \emptyset$, then:

$$\mathbb{E}[\hat{P}_k(\mathbf{x}) \mid I_k(\mathbf{x}) \neq \emptyset] = \frac{\sum\limits_{\mathbf{m} \in D_c^*} \frac{m^1}{m^0+m^1} p_D^*(\mathbf{m})}{\sum\limits_{\mathbf{m}' \in D_c^*} p_D(\mathbf{m}')}. \tag{6}$$

Notice that Equation (6) is a factorization of Equation (4), which saves a few operations in the algorithm. The algorithm for the expected proportion of successes after k consecutive successes is Algorithm 4 in Appendix R.

Going back to Example 2.3, we are now able to calculate the exact expectation of underlined flips in our hundred flip long sequence:

$$\text{exp\_prop(100, 1, 0.5)} = 0.4949495. \tag{7}$$

Due to symmetry, the algorithm for the expected proportion of successes after k consecutive failures can be easily derived from Algorithm 4. For more details see Algorithm 5 in Appendix R.

3.1.2 Expected Difference in Proportions

After establishing the procedure of calculating the expected proportions, this subsection now turns to the exact computation of the expected difference in proportions. The latter is needed to correct the bias of GVT's estimator.

The approach is very similar to the approach described in Section 3.1.1. The difference in proportions can be computed from a dictionary

$$F := (\mathbf{m} : p_F(\mathbf{m}))_{\mathbf{m} \in F_c},$$

where $F_c := \{\mathbf{m} \in \mathbb{N}^4 \mid p_F(\mathbf{m}) > 0\}$ is the set of count realizations with non-zero probability and $p_F(\mathbf{m}) := \mathbb{P}((M_0^0(\mathbf{X}), M_0^1(\mathbf{X}), M_1^0(\mathbf{X}), M_1^1(\mathbf{X})) = (m_0^0, m_0^1, m_1^0, m_1^1))$ [14]. Following the notation from 3.1.1, $M_1^0(\mathbf{X})$ and $M_1^0(\mathbf{X})$ yield the total number of failures and successes (respectively) that follow a streak of k successes. In the same way, M_0^0 and M_0^1 are the variables that yield the number of failures and successes (respectively) that follow a streak of k failures. Again the goal is to recursively define the dictionary which holds the joint probability distribution in order to calculate the expectation. Let

$$F(\ell_0, \ell_1, r)$$

[14] Since I am going into more depth than Miller and Sanjurjo (2018a), I am deviating a little from their notation here. While the just mentioned authors also denote the dictionary with which the difference in proportions is computed with D, I have opted to call it F in order to stress the difference to the dictionary D of Section 3.1.1

be the dictionary that represents the count-probability pairs for the r remaining trials of a sequence which has $\ell_0 \leq k$ successive failures and $\ell_1 \leq k$ successive successes preceding the current trial [15]. These dictionaries can be defined recursively in a similar fashion to Equation (5):

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

For a given n and k the dictionary of interest, F_c, is just dictionary $F(0, 0, n)$, i.e. the count-probability pairs for n remaining trials and zero successes and failures preceding the current trial. For the algorithm that builds the collection of dictionaries, please see Schäfer (2019) (function Count_Distribution_diff()). Due to its length, the algorithm was omitted from this thesis.

The only question that remains to be answered is how to actually calculate the expected difference in proportions, $\mathbb{E}[\hat{D}_k(\mathbf{x}) \mid I_k(\mathbf{x}) \neq \emptyset, J_k(\mathbf{x}) \neq \emptyset]$. To answer this question, it is useful to define the subset of count realizations for which the corresponding sequences fulfill $I_k(\mathbf{x}) \neq \emptyset$ and $J_k(\mathbf{x}) \neq \emptyset$, i.e.

$$F_c^* := \{\mathbf{m} \in F_c | m_0^0 + m_0^1 \neq 0 \vee m_1^0 + m_1^1 \neq 0\} \subset F_c.$$

In other words, a sequence fulfills the condition $I_k(\mathbf{x}) \neq \emptyset$ and $J_k(\mathbf{x}) \neq \emptyset$ if for the corresponding count realization of $\mathbf{x}$ holds: neither $m_0^0 = m_0^1 = 0$ nor $m_1^0 = m_1^1 = 0$. Therefore, the expected difference in proportions can now be calculated in a straightforward way:

$$\mathbb{E}[\hat{D}_k(\mathbf{x}) \mid I_k(\mathbf{x}) \neq \emptyset, J_k(\mathbf{x}) \neq \emptyset] = \frac{\sum\limits_{\mathbf{m} \in F_1^*} \frac{m_1^1}{m_1^0 + m_1^1} p_F(\mathbf{m}) - \sum\limits_{\mathbf{m}' \in F_0^*} \frac{m_0^1}{m_0^0 + m_0^1} p_F(\mathbf{m}')}{\sum\limits_{\mathbf{m}'' \in F_c^*} p_F(\mathbf{m}'')}, \tag{8}$$

where $F_0^* := \{\mathbf{m} \in F_c^* | m_0^1 \neq 0\}$ and $F_1^* := \{\mathbf{m} \in F_c^* | m_1^1 \neq 0\}$ are the subsets of F_c^* for which a success was recorded after a failure or success (respectively).

Example 3.6.
Extending Example 3.2, Table 3 reports the dictionary F_c for the case $n = 3$ and $k = 1$.

[15] Notice that $\ell_0 \ell_1 = 0$

Table 3: `Dictionary Representation of Count-Probability Pairs`

Sample space of sequences			**Dictionary** (F_c)		
Sequence	**Probability**	**Count Realization**	**Count**		**Probability**
000	q^3	$(2, 0, 0, 0)$	$\mathbf{m}$	:	$p_F(\mathbf{m})$
001	q^2p	$(1, 1, 0, 0)$	$(2, 0, 0, 0)$	:	q^3
010	q^2p	$(0, 1, 1, 0)$	$(1, 1, 0, 0)$	:	q^2p
100	q^2p	$(1, 0, 1, 0)$	$(0, 1, 1, 0)$	:	$q^2p + qp^2$
011	qp^2	$(0, 1, 0, 1)$	$(1, 0, 1, 0)$	:	q^2p
101	qp^2	$(0, 1, 1, 0)$	$(0, 1, 0, 1)$	:	qp^2
110	qp^2	$(0, 0, 1, 1)$	$(0, 0, 1, 1)$	:	qp^2
111	p^3	$(0, 0, 0, 2)$	$(0, 0, 0, 2)$	:	p^3

Notes: In the table to the left Columns 1 and 2 are the same as Columns 1 and 2 from Table 2. Column 3 displays the corresponding count realizations, while the first two entries report the number of failures and successes after a failure and the latter two the number of failures and successes after a success. The table to the right reports the dictionary. Column 1 represents the list of keys (unique count realizations) and Column 2 the list of values (probabilities of the count realizations).

It is now easy to see that:

$$\begin{aligned} F_c^* &= \{(0, 1, 1, 0), (1, 0, 1, 0), (0, 1, 0, 1)\}, \\ F_0^* &= \{(0, 1, 1, 0), (0, 1, 0, 1)\}, \text{ and} \\ F_1^* &= \{(0, 1, 0, 1)\}. \end{aligned}$$

This gives us:

$$\begin{aligned} \mathbb{E}[\hat{D}_1(\mathbf{x}) \mid I_1(\mathbf{x}) \neq \emptyset, J_1(\mathbf{x}) \neq \emptyset] &= \frac{qp^2 - ((q^2p + qp^2) + qp^2)}{2qp^2 + 2q^2p} = \frac{-(q^2p + qp^2)}{2qp(p + q)} \\ &= \frac{-(q + p)}{2(p + q)} = -\frac{1}{2}. \end{aligned}$$

This means for $n = 3$, $k = 1$ and any given probability of success p the expected difference in proportions is -0.5.

3.2 Correcting the Bias of GVT's Estimator

This subsection will de-bias the estimator and redo the tests performed by GVT. As Miller and Sanjurjo (2018b) did, this thesis will focus on the third (controlled) study for the bias correction. Without any further controls for in-game factors, the other two studies performed by GVT are unsuitable for the study of hot hand shooting [16]. Furthermore, the case of $k = 1$ will also be disregarded in this thesis as it is vulnerable to a measurement error problem, which was first pointed out by Stone (2012). To keep it short, the act of making a single shot is a weak signal for the change of a player's underlying state (normal state to the hot state) and therefore not qualified to be a subject of hot hand analysis. Consequently, the longer the streak the stronger the signal for the hot hand. For more details see Miller and Sanjurjo (2018b, footnotes 20 and 22) and Stone (2012). Miller and Sanjurjo (2018b) put the focus in their work specifically on the case $k = 3$ and justified this with a reference to the work of Carlson and Shu (2007), who show that people typically perceive a streak as beginning with the third successive event. They do, however, also correct the bias and redo the test for the cases $k = 2$ and $k = 4$ (a "stronger signal of hot hand shooting"), yet the discussion of the obtained results was reduced to a footnote. This thesis will go into more detail regarding the math of the performed tests and corresponding standard errors as well as regarding the cases $k = 2$, $k = 4$[17]. I will uncover why Miller and Sanjurjo (2018b) might have had a compelling reason to keep the discussion of the latter short and also point out a small mistake they made in a calculation of a p-value.

3.2.1 k = 3

Table 4 replicates in part Table 4 of GVT and also in part Table II of Miller and Sanjurjo (2018b). Columns 2 to 4 reproduce the shooting performances of each player that appear in Columns 2, 5, and 8 of Table 4 of GVT, using the raw data. Contrary to GVT and Miller and Sanjurjo (2018b), I have rounded to the nearest thousandth to highlight the differences (or equalities) in the percentages a little more. The number of shots in each category is put in brackets. GVT's estimator can be found in Column 5, which is just the difference of Columns 2 and 4. Notice that player 12 of the females was excluded (in this column and further analysis) since there was no shot recorded in the category ''3 hits". To test for the hot hand, GVT performed a paired t-test on Column 5. The null hypothesis was $\mathbb{E}[\hat{D}_3^i] = 0$, i.e. there is no hot hand. The performed t-test failed to reject the null hypothesis ($p = 0.49$) (see Schäfer (2019) for more [18]). However, Theorems 2.6 and 2.7 clearly point out the flaw in the null hypothesis and therefore make the result of the t-test obsolete. Miller and Sanjurjo (2018b) propose two alternatives to redo the tests with a more sensible null hypothesis.

The first step is correcting the GVT estimate (Column 5) for the corresponding bias. Column 6 of Table 4 calculates the expected difference in proportions for each player taking the number of shots and baseline hit rate (Column 3) into account. This was done by using Algorithm

[16] for more details see Section 4 and Miller and Sanjurjo (2018b, footnotes 21 and 22).

[17] Although Miller and Sanjurjo (2018b) disregard the case of $k = 2$ a little bit due to the above mentioned findings of Carlson and Shu (2007), I disagree with this reasoning. The perception of a streak by people has nothing to do with a player's underlying probability of success (taking on the hot hand state). Although shorter streak lengths exacerbate the attenuation bias due to measurement error, Stone (2012) only mentioned the bias to be a significant problem for the case $k = 1$.

[18] All the tests that are discussed in the following can be found in Schäfer (2019)

6, which was explained in Section 3. For example, the first entry of Column 6 was calculated through exp_diff(100, 3, 0.54) (= -0.081). Column 7 reports the bias adjusted estimate,

$$\hat{A}_3 := \hat{D}_3 - \text{bias}_3,$$

which is just the GVT estimate subtracted by the bias.

Table 4: `Shooting Performance and Adjusting the Bias for` $k = 3$

shooter	$\hat{P}$(hit|3 makes)	$\hat{P}$(hit)	$\hat{P}$(hit|3 misses)	GVT est. ($\hat{D}_3$)	bias$_3$	bias adj. ($\hat{A}_3$)
Males						
1	0.5 (12)	0.54 (100)	0.444 (9)	0.056	-0.081	0.137
2	0 (3)	0.35 (100)	0.429 (28)	-0.429	-0.099	-0.330
3	0.6 (25)	0.6 (100)	0.667 (6)	-0.067	-0.090	0.023
4	0.333 (3)	0.4 (90)	0.467 (15)	-0.133	-0.106	-0.027
5	0.333 (6)	0.42 (100)	0.75 (12)	-0.417	-0.086	-0.330
6	0.652 (23)	0.57 (100)	0.25 (12)	0.402	-0.085	0.487
7	0.647 (17)	0.56 (75)	0.286 (7)	0.361	-0.115	0.476
8	0.571 (7)	0.5 (50)	0.5 (6)	0.071	-0.137	0.209
9	0.833 (30)	0.54 (100)	0.35 (20)	0.483	-0.081	0.565
10	0.571 (21)	0.6 (100)	0.571 (7)	0.000	-0.090	0.090
11	0.619 (21)	0.58 (100)	0.571 (7)	0.048	-0.086	0.134
12	0.429 (7)	0.44 (100)	0.412 (17)	0.017	-0.083	0.100
13	0.5 (18)	0.61 (100)	0.4 (5)	0.100	-0.092	0.192
14	0.6 (20)	0.59 (100)	0.5 (6)	0.100	-0.088	0.188
Females						
1	0.333 (9)	0.48 (100)	0.667 (9)	-0.333	-0.080	-0.253
2	0.4 (5)	0.34 (100)	0.429 (28)	-0.029	-0.101	0.072
3	0.5 (8)	0.39 (100)	0.36 (25)	0.140	-0.092	0.232
4	0.333 (3)	0.32 (100)	0.267 (30)	0.067	-0.103	0.170
5	0.2 (5)	0.36 (100)	0.222 (27)	-0.022	-0.097	0.075
6	0.286 (7)	0.46 (100)	0.545 (11)	-0.260	-0.081	-0.179
7	0.615 (13)	0.41 (100)	0.32 (25)	0.295	-0.088	0.383
8	0.733 (15)	0.53 (100)	0.667 (9)	0.067	-0.080	0.147
9	0.5 (8)	0.45 (100)	0.462 (13)	0.038	-0.082	0.121
10	0.714 (14)	0.46 (100)	0.316 (19)	0.398	-0.081	0.480
11	0.385 (13)	0.53 (100)	0.5 (10)	-0.115	-0.080	-0.035
12	- (0)	0.25 (100)	0.324 (37)			
Average	0.487	0.472	0.454	0.034	-0.092	0.126

Notes: Columns 2, 3, and 4 reproduce Columns 8, 5, and 2 of GVT's Table 4 (respectively) using the raw data. Column 5 displays GVT's estimate, i.e. the difference between Column 2 and 4. Column 6 reports the exact bias for this estimate using Algorithm 6 (see Appendix R). The bias adjusted difference (Column 5 - Column 6) is listed in Column 7.

After the correction, we can once again perform a t-test under the null hypothesis that the trials are independent Bernoulli-distributed random variables and check if the bias adjusted difference is significantly different from zero. Indeed, with a mean of 0.126, the t-test reveals that the bias adjusted difference is significantly greater than zero ($p = 0.015$), thus giving us evidence for the existence of the hot hand.
Notwithstanding, the just performed t-test suffers from one significant drawback. The bias adjusted difference does not account for the amount of shots taken in each category (''3 hits'' and ''3 misses'') and therefore reduces the player's shooting performance to a single number. For example, in the calculation of the standard error (which in turn is needed to calculate the p/T-values), the adjusted difference of player 2 on the male side (who took only 3 shots in the category ''3 hits'' but a whopping 28 in the category ''3 misses'') is equally taken into account as the difference of, for example, player 8 for the males, who took a rather equal amount of shots in each category (7 and 6 respectively). For more details see Appendix S.
As a result, Miller and Sanjurjo (2018b) propose a different method of calculating the standard error and p-value of the statistical model, a method where the shots each player took in each of the categories is taken into account. They obtain p $<$ 0.01 and S.E. $=$ 4.7 percentage points. Since Miller and Sanjurjo (2018b) provide no details on the calculation in their paper and the numbers seem to appear out of nowhere, I will show and explain the math that was done here. For their approach the just mentioned authors only perform a one-sided test (H_1: $\mathbb{E}[\hat{A}_3^i] = \mathbb{E}[\hat{D}_3^i - \text{bias}_3^i] > 0$), which is of course sufficient since we are only testing if the performance of players sometimes exceeds expectation. Next, Miller and Sanjurjo (2018b, Footnote 25) compute the standard error based on the assumption of independence across the 2515 trials and normality. I performed a Shapiro-Wilk test to confirm that the latter assumption is sensible. The null hypothesis (the data is normally distributed) failed to be rejected. The p-value of such a test can now be calculated as:

$$p = 1 - \text{pnorm}\left(\frac{\bar{A}_k}{SE(\bar{A}_k)}\right), \tag{9}$$

where pnorm is the probability density function of the standard normal distribution, and $\bar{A}_k := \frac{1}{n}\sum_{i=1}^{n} \hat{A}_k^i$ the mean bias-adjusted difference, which is used as the estimator for $\mathbb{E}[\hat{A}_k^i]$. Finally, $SE(\bar{A}_k)$ denotes the standard error of $\bar{A}_k$. Calculating the latter is a little tricky. The standard error is an estimate of the standard deviation of the sampling distribution. Therefore:

$$SE(\bar{A}_k) = \sqrt{\widehat{\text{Var}}(\bar{A}_k)} = \sqrt{\frac{1}{n^2}\sum_{i=1}^{n} \widehat{Var}(\hat{D}_k^i)}. \tag{10}$$

For the second equality, the independence as well as basic variance properties are used. Miller and Sanjurjo (2018b) now proceed to estimate the variance as

$$\widehat{\mathrm{Var}}(\hat{D}_k^i) = (SE_{pooled_k}^i)^2. \tag{11}$$

$SE_{pooled_k}^i$ is the pooled standard error, i.e. the standard error of a bias-corrected player-by-player two-sample t-test assuming equal variances. The pooled standard errors are computed as follows: Using GVT's raw data, the variance of the sequence of shots which follow three successive hits and the variance of the sequence of shots which follow three consecutive misses is computed. Then the pooled standard deviation can be computed as

$$SD_{pooled} = \sqrt{\frac{(n_1 - 1) * \mathrm{Var}_1 + (n_2 - 1) * \mathrm{Var}_2}{n_1 + n_2 - 2}}, \tag{12}$$

where n_1 and n_2 are the number of shots, and Var_1 and Var_2 are the variances of the binary shot sequences of the categories "3 hits", "3 misses" respectively. The pooled standard error for player i is then given as

$$SE_{pooled_k}^i = SD_{pooled_k}^i * \sqrt{\frac{1}{n_1^i} + \frac{1}{n_2^i}}. \tag{13}$$

I will demonstrate the calculation of the pooled standard error with player 10 of the males. Player 10 shot 21 shots in the category "3 hits" ($\Rightarrow n_1 = 21$) and 7 shots in the category "3 misses" ($\Rightarrow n_2 = 7$). Looking at the raw data, player 10 made 12 of his 21 shots in the first category and 4 of his 7 in the latter. Therefore, the shot sequences in both categories could look like $X_1 = 111111111111000000000$ or $X_2 = 1111000$ respectively [19]. Now,

$$\mathrm{Var}_1 = \mathrm{Var}(X_1) = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (x_{1i} - \overline{X}_1)^2 \approx 0.257.$$

x_{1i} denotes the ith realization of the sequence X_1. To ensure consistency, the sum is divided by $n_1 - 1$ rather than n_1. In the same way one calculates $\mathrm{Var}_2 \approx 0.286$. Now we possess the necessary tools to calculate the pooled standard deviation and pooled standard error as demonstrated in Equations 12 and 13. Thus, for player 10 we obtain

$$SD_{pooled_3}^1 0 \approx 0.514 \text{ and } SE_{pooled_3}^1 0 \approx 0.224.$$

These calculations can be done accordingly for each of the 25 shooters[20](see Schäfer (2019)). The variance of the average difference across players is now given by $\widehat{\mathrm{Var}}(\overline{D}_3)$, where $\overline{D}_3 =$

[19] For the calculation of the variance it does not matter in which order the makes and the misses came. A representation like this, however, is easier to code.

[20] Remember, player 12 of the females was excluded.

$\frac{1}{n}\sum_{i=1}^{n}\hat{D}_3^i$ and $\hat{D}_3^i$ is player i's difference, $\hat{D}_3^i := \hat{P}^i(\text{hit} \mid 3 \text{ hits}) - \hat{P}^i(\text{hit} \mid 3 \text{ misses})$. Therefore,

$$\widehat{\text{Var}}(\overline{D}_3) = \frac{1}{n^2}\sum_{i=1}^{n}\widehat{\text{Var}}(\hat{D}_3^i) = \frac{1}{25^2}\sum_{i=1}^{n}(SE^i_{pooled_3})^2) \approx 0.002,$$

and

$$SE(\overline{A}_3) = \sqrt{\widehat{\text{Var}}(\overline{D}_3)} \approx 0.0466 \approx 4.7 \text{ percentage points.}$$

Due to normality, the one-sided test has a p-value of

$$p = 1 - \text{pnorm}\left(\overline{A}_3/SE(\overline{A}_3)\right) \approx 0.003.$$

While GVT argued that most of the players in their controlled study shot relatively better after a streak of misses than after a streak of hits (see for example Column 5 of Table 4, where 14 of the 25 players have a negative sign.), Miller and Sanjurjo (2018b) show that after the bias correction 19 of the 25 players actually have a positive sign for the difference. Figure 3 replicates Figure 2 of Miller and Sanjurjo (2018b), and shows the bias-corrected difference $\hat{A}_3$ for each player in ascending order. Along with that, the 95 % confidence intervals and standard errors are shown. The standard errors are computed as shown above and the confidence intervals are then computed straightforwardly ($\hat{A}^i_k \pm z_{\alpha/2} * SE^i_{\text{pooled}_k}$).

Miller and Sanjurjo (2018b) now performed a t-test to check how many players exhibit significant hot hand shooting. They obtained a number of 5 players (p < .05, t-test, see Schäfer (2019) for the calculation). We can confirm this by looking closely at Figure 3. We can see that for 5 of the 25 players the confidence intervals are strictly greater than 0. This itself is significant as a binomial test performed on this data reveals (p < 0.01) [21].

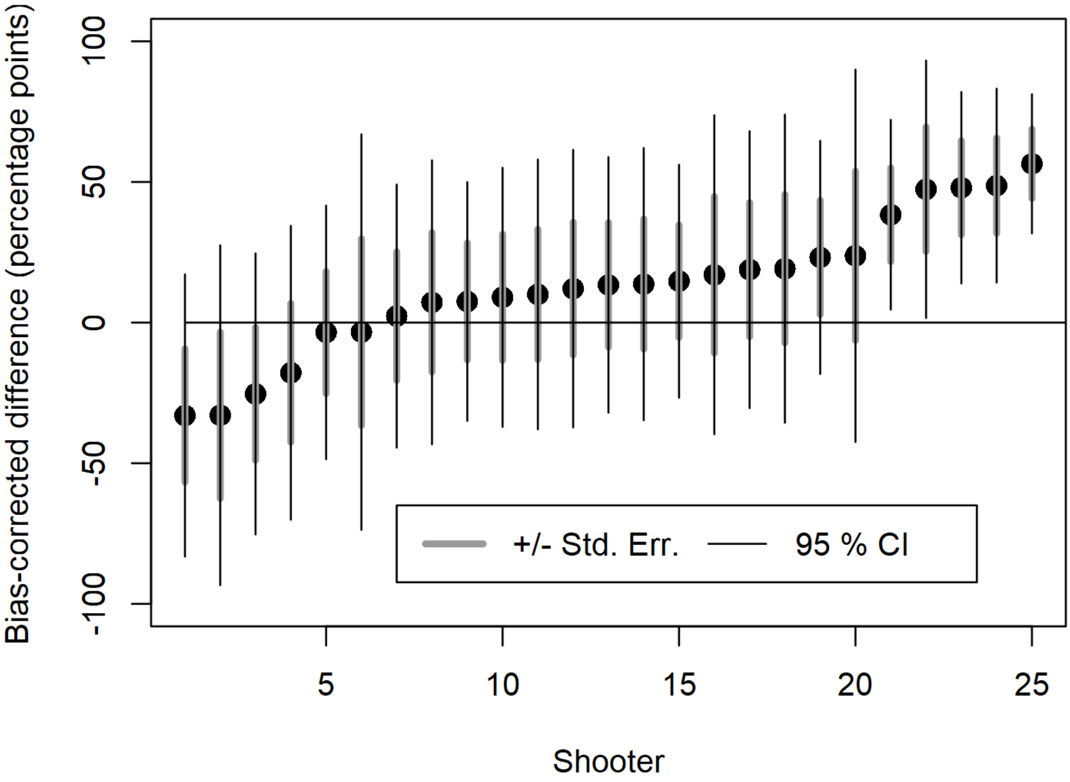
It has become evident that after the bias-correction, the case $k = 3$ yields strong evidence for the existence of the hot-hand. Let us see if the cases $k = 2$ and $k = 4$ yield similar results.

3.2.2 k = 2, 4

Tables 8 and 9 in Appendix T report the same results for the streak lengths 2 and 4 as did Table 4 for the case of $k = 3$. Since the construction of those two tables is exactly the same as Table 4, they were outsourced to the appendix to ensure a better flow in the main part of the thesis. Moreover, the same tests were repeated for the cases $k = 2$ and $k = 4$. Since the procedure of the tests was explained at length, above I will only summarize the results. While the paired t-test on the GVT-estimate (Column 5) unsurprisingly does not reject the null hypothesis (there is no hot hand), it is surprising that in both cases the t-test on the bias-corrected estimates (Column 7) do not reject the null hypothesis either. In both cases the

[21] The binomial test has the following structure. Number of trials: 25. Number of successes: 5. Probability of a success: 0.05. H_0 : the probability of success is 0.05, H_1 : the probability of success is greater than 0.05.

Figure 3: `Bias-Corrected Difference for k = 3`



bias-adjusted difference is not significantly different from zero ($p = 0.14$ and 0.23 respectively). This is something Miller and Sanjurjo (2018b) conveniently left out in the documentation of their work. Nonetheless, luckily the player-by-player tests, where the standard errors take the number of shots each player shot in each category into account, help to strengthen their case for the existence of the hot hand. As the bottom rows of Tables 8 and 9 display, the average bias-adjusted difference in proportions are $\bar{A}_2 = 0.053$ and $\bar{A}_4 = .102$. The one-sided test, $H_0 : \mathbb{E}[\hat{A}^i_k] = 0$ and $H_1 : \mathbb{E}[\hat{A}^i_k] > 0$ reveals $p = 0.02$, $SE = 6.9$pp for $k = 4$ and $p = 0.04$, $SE = 3$pp for $k = 2$, thus confirming that the findings for $k = 3$ were no fluke.

For the case $k = 4$ Miller and Sanjurjo (2018b) actually report a p-value of 0.07. Looking through their code and the circumstances that led to this result, I believe they made a minor mistake in their calculation. I will show their mistake and also show that upon modifying the calculation the p-value is actually 0.02, as stated above. Looking at Table 9 one can see that player 2 of the males and player 12 of the females were excluded from the analysis since no shot for both of them was recorded in the category "3 hits". Therefore, the average of Columns 5, 6, and 7 is taken over 24 players instead of 26. There is one more crucial thing to notice. Shooters 4 and 5 of the females as well as shooter 4 of the males each only shot one shot in the category "4 makes". Although this makes them eligible for the difference in means analysis, the standard error for the two-sample t-test with equal variances ($SE^i_{pooled_4}$) cannot be computed as one can infer from Equations 9, 10, 11, 13, and 12 [22]. Therefore, the standard error of 6.9 percentage points was calculated without those observations. As a result, if we want to compute the p-value as described in Equation 9, we need to exclude those observations from the calculation. Here is where Miller and Sanjurjo (2018b) made a mistake. Their calculation was:

$$p = 1 - \text{pnorm}\left(\frac{0.102}{0.069}\right) \approx 0.07.$$

Notice that the mean in the nominator of the fraction is taken over 24 players while the standard error on the other hand was only calculated using 21 players (Males' player 4 and Females' players 4 and 5 were excluded as just explained). The mean over the 21 players included in this analysis is actually 0.149. This leads us to a p-value of

$$p = 1 - \text{pnorm}\left(\frac{0.149}{0.069}\right) \approx 0.015.$$

Analogously to Figure 3, Figures 6 and 7 in Appendix G show the ascending bias-corrected differences in percentage points with their standard errors and confidence intervals for the cases $k = 2$ and $k = 4$ respectively.

Figure 6 shows that 14 of the 26 shooters report a positive bias-corrected difference while 4 of which even exhibit significant hot-hand shooting (p $<$ .05, t-test) [23]. This itself is significant as a one sided binomial test in this case reveals ($p = 0.04$).

Notice that there are 4 dots without an associated standard error or confidence interval in Figure 7 . The first dot corresponds to player 5 of the males, who took two shots in the category "4 makes" but failed to convert either of them and at the same time converted all of the three shots he took in the category "4 makes". Consequently, the standard error of the two-sample t-test under the equal variances assumption is zero, resulting in a confidence interval corresponding to just a single point, the bias-adjusted difference [24]. The other "naked" dots correspond to the shooter 4 of the males as well as shooters 4 and 5 of the females, as the standard errors cannot be calculated for those observations. Nevertheless, 15 of the 24 players show a bias-corrected difference of greater than zero. Although only three confidence intervals are strictly greater than zero, the t-test reveals that the sixth shooter from the left also exhibits significant hot hand shooting ($p = 0.03$). Overall, 4 of the 21 eligible players for this analysis exhibit significant hot hand shooting (p $<$ 0.05, t-test). A binomial test on this result discloses that this itself is also significant ($p = 0.02$).

[22] The key observation is that the variance of a sequence of length 1 cannot be determined.

[23] This is again confirmed by the confidence intervals, as four confidence intervals are strictly greater than zero (the lower bound of the confidence interval of the shooter fourth from the left, starts at $0.003 > 0$).

[24] The key observation here is that the variance of both shot sequences is zero, resulting in a pooled standard deviation of zero (see Equations 12 and 13).

This subsection made it evident that after correcting for the bias, the interpretation of the results of GVT seem to reverse and yield strong evidence for the existence of the hot-hand.

4 The Hot Hand Effect on the Team

While most studies so far have solely focused on the hot hand effect of an individual, the application part of this thesis will provide an approach to an examination if the hot hand effect is detectable on a whole team. I use a dataset provided by BigDataBall. This dataset consists of all the play-by-play data of each game of the 2017-2018 NBA season. As GVT mention in their paper concerning studies one and two (see Section 2.1), one very significant drawback of analyzing in-game data is the fact that there are several factors that come into play that might mask the hot hand effect. Professional players who enter the hot state typically feel more confident in their ability to shoot and have reported feeling that they "almost can't miss" (GVT, p. 302). As a result, they may try to shoot from a further distance more frequently than they normally would. Moreover, the opposing team might try to guard players who "get hot" more closely than before or let their best defender guard the player in the hot state [25]. Rao (2009) and Bocskocsky et al. (2014) show that shot difficulty, in fact, tends to increase following several made shots, invalidating the assumption of shot selection independence. Therefore, I created a model which controls (at least partially) for shot difficulty in my analysis. It will be explained in detail in Section 4.3.

The crucial advantage of an analysis of the hot hand effect on a team level compared to the individual level is that the analyzed shots are linked more closely to one another as the time difference between two shots is often less than a minute. This ensures a more promising signal of the hot state (of a team). When analyzing the hot hand at the individual level, all the papers I found used only one premise of a valid shot sequence. This premise is that the shots of the analyzed sequence came from the same game (see for example Bocskocsky et al. (2014) and Rao (2009)). This approach disregards the time that lies in between shots. For example, a certain player could have shot three for three from the field in the first quarter and then not shot the ball until the middle of the fourth quarter. By GVT's definition, this player was in the "hot-state" when he was shooting the shot in the fourth quarter even though it is reasonable to assume that an elevation of the player's ability to make this shot does not necessarily have to extend over three quarters. There are a number of factors that can influence a player's general confidence. For example, a missed defensive assignment could lead to a harsh reprimand by the coach, which in turn could lead to a decrease of confidence. Another example could be that the coach does not give the player enough playing time in the quarters two and three as the player feels like he deserves. Some players might also argue that an intermission, like getting subbed out for two minutes, is enough time to break ones' "flow".

[25] This particular effect, although probably mildly, also translates to the team level, as Bocskocsky et al. (2014) for example have shown that a player in the hot state is more likely to take the next shot than he normally would.

These are just a few examples of possibly thousands that can influence or even completely change a player's state of mind. To be clear, I am not saying that the increased time difference of shots in analyses at the individual level is a problem, or that a player's state of mind cannot also change in a relatively short period of time. However, I think it is undeniable that shots that are closely linked in time are more comparable with regard to an analysis of the hot hand effect than shots that may have more than half a game between them. Therefore, a streak of makes might be an even better signal for the hot hand state at the team level than it is at the individual level.

Due to reasons I will extensively explain below, this application part is merely supposed to give a rough idea on how one could proceed to test for a hot hand on the team level. For an extensive analysis I unfortunately lack data (see Section 4.2 for more). Therefore, the statistical power of my results (see Sections 4.5 and 4.6) is limited.

In Section 4.1 a quick overview of the used dataset is given. Section 4.2 deals with the implementation of GVT's model on the data. A model, which takes the shot difficulty into account is then the subject of Sections 4.3 - 4.6. An outlook of possible extensions of the model is discussed in section 4.7.

4.1 Raw Data

The dataset used in this analysis was obtained from the website `bigdataball.com`, which provides historical NBA play-by-play game logs since the 2004-2005 season [26]. The specific dataset in question is the play-by-play game logs from the 2017-2018 NBA season (including playoffs). It includes all the plays that happened during that season. Since the analysis of this thesis concerns only the plays where a shot from the field was involved, the dataset was first trimmed and adjusted for my purposes. After cleaning, the dataset included all the 225,490 shots that were taken from the field during the 2017-18 season along with details about the shot circumstances (see Schäfer (2019) and Section 4.3).

4.2 GVT's Approach on the Team Level

Leaning on the analysis of Section 3.2, Table 5 provides an analogue to Tables 4, 8, and 9 for our dataset. Column 1 reports the team in question, and the other columns the already familiar probabilities and estimators while the number of shots is again put in brackets. As mentioned earlier, GVT's estimator is actually a consistent estimator and the bias virtually vanishes for a trial length of several thousands shots. Looking at Column 4, we see that every team has shot at least 6999 shots throughout the 2017-18 season. This means, GVT's estimator $\hat{D}_k^i$ is not vulnerable to the selection bias in this case (i stands for one of the 30 teams in this section). Columns 7 and 8 report the estimator for streak lengths 3 and 4 respectively. As Definition 2.1 states, the greater the estimator, the greater the implication for the existence

[26] A play-by-play game log is a transcript of all the plays/actions that happen during a game.

Table 5: Probability of Making a Shot Conditioned on the Outcome of Previous Shots for All NBA Teams

Team	$\hat{P}$(hit|4 misses)	$\hat{P}$(hit|3 misses)	$\hat{P}$(hit)	$\hat{P}$(hit|3 makes)	$\hat{P}$(hit|4 makes)	GVT est. k = 3 ($\hat{D}_3$)	GVT est. k = 4 ($\hat{D}_4$)
LAL	0.495 (535)	0.481 (1044)	0.461 (7248)	0.422 (635)	0.43 (265)	-0.059	-0.065
LAC	0.483 (501)	0.481 (969)	0.472 (6999)	0.466 (731)	0.451 (337)	-0.014	-0.032
OKC	0.447 (689)	0.444 (1255)	0.451 (7739)	0.463 (722)	0.45 (331)	0.019	0.003
NYK	0.445 (562)	0.459 (1058)	0.464 (7190)	0.442 (670)	0.411 (292)	-0.018	-0.034
TOR	0.497 (537)	0.488 (1073)	0.472 (7991)	0.438 (755)	0.415 (328)	-0.050	-0.083
CHI	0.442 (702)	0.441 (1268)	0.435 (7284)	0.425 (584)	0.425 (247)	-0.016	-0.016
SAC	0.409 (618)	0.444 (1121)	0.45 (7061)	0.475 (634)	0.439 (301)	0.031	0.029
PHX	0.43 (683)	0.427 (1209)	0.442 (7140)	0.45 (596)	0.487 (265)	0.023	0.056
POR	0.474 (606)	0.476 (1160)	0.452 (7496)	0.446 (624)	0.434 (274)	-0.030	-0.039
MIN	0.503 (479)	0.509 (987)	0.476 (7475)	0.461 (722)	0.459 (331)	-0.047	-0.044
SAS	0.468 (583)	0.471 (1122)	0.455 (7417)	0.441 (682)	0.432 (301)	-0.029	-0.036
DEN	0.463 (562)	0.454 (1040)	0.47 (7102)	0.472 (724)	0.46 (337)	0.019	-0.003
UTA	0.475 (594)	0.471 (1144)	0.461 (7706)	0.461 (725)	0.483 (329)	-0.010	0.009
DAL	0.453 (647)	0.439 (1168)	0.444 (7040)	0.454 (582)	0.459 (259)	0.014	0.007
ATL	0.467 (599)	0.455 (1116)	0.446 (7016)	0.428 (563)	0.462 (238)	-0.027	-0.005
NOP	0.508 (512)	0.5 (1036)	0.483 (8048)	0.494 (866)	0.47 (423)	-0.006	-0.037
MEM	0.428 (656)	0.415 (1141)	0.444 (6788)	0.439 (590)	0.395 (256)	0.024	-0.034
MIL	0.449 (492)	0.486 (978)	0.48 (7355)	0.491 (782)	0.485 (377)	0.005	0.036
PHI	0.466 (575)	0.472 (1114)	0.468 (7983)	0.459 (767)	0.42 (348)	-0.013	-0.047
WAS	0.491 (552)	0.485 (1086)	0.467 (7525)	0.488 (722)	0.479 (351)	0.002	-0.012
ORL	0.457 (647)	0.44 (1168)	0.452 (7038)	0.479 (639)	0.521 (303)	0.039	0.064
MIA	0.439 (649)	0.445 (1184)	0.454 (7411)	0.429 (655)	0.462 (277)	-0.016	0.023
BKN	0.445 (638)	0.443 (1169)	0.441 (7112)	0.41 (586)	0.39 (236)	-0.034	-0.055
IND	0.499 (565)	0.471 (1082)	0.473 (7642)	0.479 (778)	0.48 (369)	0.008	-0.019
DET	0.434 (640)	0.428 (1137)	0.45 (7128)	0.458 (618)	0.451 (277)	0.030	0.017
CHA	0.423 (615)	0.45 (1147)	0.45 (7105)	0.469 (646)	0.43 (302)	0.019	0.008
GSW	0.48 (508)	0.505 (1044)	0.498 (8786)	0.49 (1044)	0.496 (508)	-0.014	0.016
HOU	0.465 (660)	0.468 (1262)	0.456 (8353)	0.453 (762)	0.446 (341)	-0.015	-0.019
CLE	0.509 (625)	0.491 (1240)	0.471 (8756)	0.438 (880)	0.446 (379)	-0.054	-0.063
BOS	0.458 (757)	0.451 (1402)	0.448 (8556)	0.449 (719)	0.416 (322)	-0.002	-0.042
Average	0.463	0.463	0.46	0.456	0.45	-0.007	-0.014

of the hot hand. It neither takes a trained eye nor a paired t-test to see that the results are not in favor of the hot hand. If anything, they would rather confirm the existence of the hot hand fallacy as 18 of the 30 teams have a negative sign in front of the estimator for $k = 3$ and 19 of 30 for $k = 4$. Nevertheless, as already mentioned several times above, for in-game data GVT's estimator very likely suffers from an omitted variable bias. In games, no shot is truly comparable to the other as distance, fatigue, defensive pressure etc. vary. To account for this we need a model that incorporates shot difficulty. This is the subject of the next subsection.

4.3 Predicted Shot Difficulty

Using the shot log data described in Section 4.1, I estimate a model that predicts the shot difficulty of each shot for team i taking shot s. The approach is inspired by Bocskocsky et al. (2014). The shot difficulty is predicted based on three main categories of determinants of shot difficulty.

$$\hat{P}_{is} = \alpha + \beta \cdot (\text{Game Condition Controls}_{is}) + \gamma \cdot (\text{Shot Controls}_{is}) + \delta \cdot (\text{Team Fixed Effects}_{i})$$

To ease notation, throughout this section $\hat{P}$ will denote the estimated probability of a make. The *game condition controls* are variables like `period, score_diff, remaining_time` (in the quarter), and `play length`. These variables are supposed to proxy for differences in player fatigue, defensive pressure and general effort. *Shot controls* include features like `shot_distance`, `shot_type` and `shot_adddiff`. Lastly, the Team Fixed Effects is supposed to control for the difference in abilities that different teams in the NBA have. For example, it makes little sense to compare the Golden State Warriors, which had four all-stars on their team during the 2017-2018 season with, for instance, the Phoenix Suns, which only managed to win around 26 % of their games during that season.[27]. All the variables included in the model can be observed in Table 10 in Appendix T.

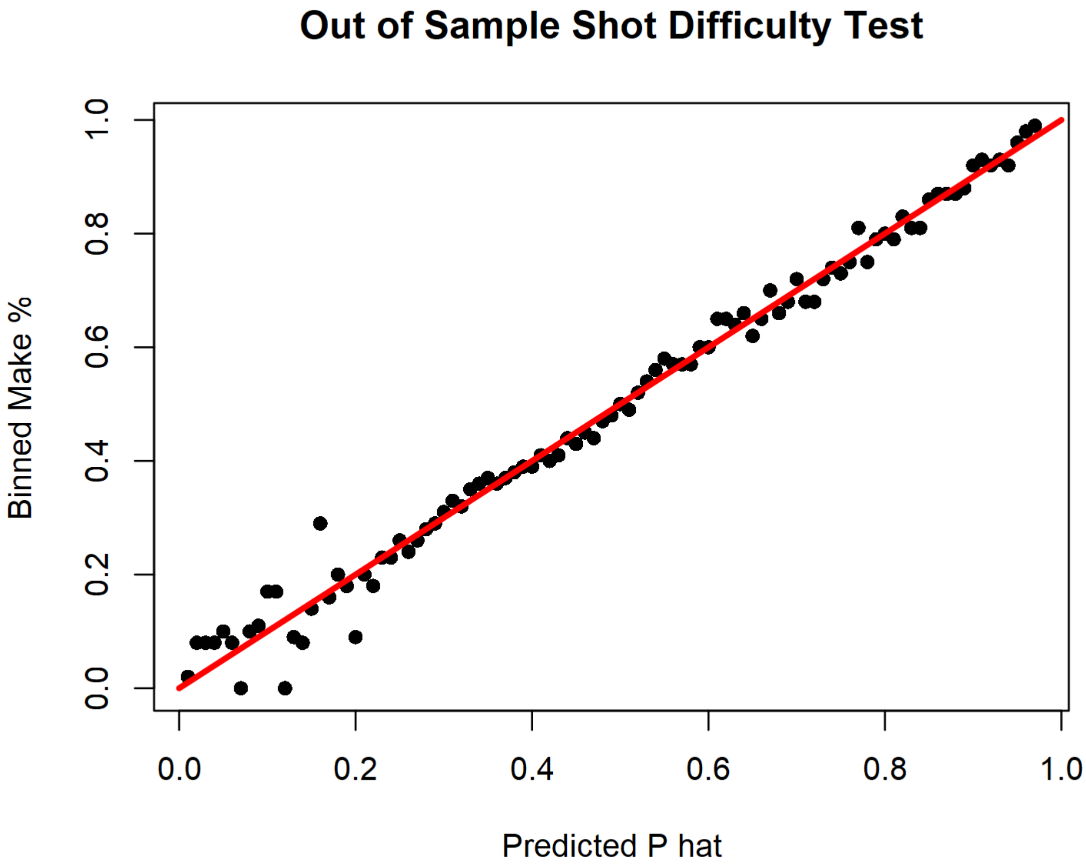
One major difference between this thesis' model and the model of Bocskocsky et al. (2014) is the absence of a *defensive control*-category as a determinant for shot difficulty. Unfortunately, I did not have access to any data that includes defensive controls, including, but not limited to, the distance between the player shooting and the closest defender, the angle of that defender relative to a straight line between the shooter and the basket, and the height difference between shooter and defender. Such data is collected by an optical tracking system using special cameras and is generally not available to the public. My model is therefore vulnerable to an omitted variable bias. However, as Bocskocsky et al. (2014) point out, even highly sophisticated shot difficulty models can be vulnerable to that bias. In Section 4.7 I will discuss possible extension of my model for individuals with richer datasets so that the risk of an omitted variable bias can be reduced.

[27] Every season only 24 players get selected to play in the all-star game. Having four of them on one team is a rare exception.

4.3.1 Regression Results

The model was fit using a Logit model. If the model predicted a value of one or zero, I replaced them with 0.99 and 0.01 respectively as no shot in basketball has a 100 % or 0 % chance of going in. Due to the large number of variables considered in the model, I decided to forego a report of the over 50 coefficients. However, the code for the regression fit is available for reproduction on GitHub (see Schäfer (2019)) [28]. Similarly to Bocskocsky et al. (2014), the accuracy of the model was tested by running the model on a randomized training set consisting of half of the games. The other half of the games was used as a validation set for which the $\hat{P}$'s were then predicted. The probabilities were rounded to nearest percent, grouping the $\hat{P}$'s into bins of one percent increments. For each bin, the actual make-percentage was then calculated. The more accurate the model, the more closely each bin should correspond to the actual make percentages. Figure 4 shows the scatter plot of the data. One can see, the model fits the data quite well.

Figure 4: `Bias-Corrected Difference for k = 4`



In the following analysis, $\hat{P}$ is often used as a dependent or independent variable of models. It goes without saying that for the fitting of those models only the realizations in the validation

[28] The data itself is not freely available, and therefore unfortunately not available on GitHub.

set are used [29]. In addition, if in the following is spoken about the "actual probability of a make", then this refers to the binned make percentages described above, and visible in Figure 4.

4.4 Definition of Heat of a Team

As already mentioned in Section 2, in order to test for the hot hand, an operational definition is needed. In this analysis a definition that incorporates shot difficulty is necessary. I will use the definition provided by Bocskocsky et al. (2014), which they refer to as *Complex Heat*$_n$. For simplicity I will just call it *Heat*$_n$. It is defined as:

$$\mathit{Heat}_n = \text{Actual \% over past } n \text{ shots} - \text{Expected \% over past } n \text{ shots.}^{30} \tag{14}$$

This means, the greater *heat*, the "hotter" the team. I simulated the analysis in the following subsections for $n = 3$ through $n = 6$ and obtained similar results. Bocskocsky et al. (2014) found $n = 4$ to be a reasonable number of shots, so I will also use $n = 4$ for the report of results of my analysis as well.

Using the above definition of heat enables us to account for shot difficulty. While GVT's definition would label a shooter shooting 4 layups in a row just as "hot" as a shooter shooting three three-pointers in a row, Equation (14) corrects for this and would label the shooter shooting the threes as "hotter".

4.5 The Connection Between Heat and Shot Difficulty

It is logical that hot players/teams would tend to take more difficult shots since "hotness" is characterized by a temporary elevation of ability to make a shot, i.e. players feel comfortable taking shots they would pass up under normal circumstances. Subsequently, before actually testing for the hot hand, it is natural to first check the effect of heat on shot difficulty. Therefore, I run

$$\hat{P}_{is} = \alpha + \beta * (Heat_{is}),$$

using a regular OLS-regression. If increasing heat would lead players to take more difficult shots, we would expect β to be negative (recall that a high $\hat{P}$ corresponds to a relatively easier shot).

Table 6 shows the result of the above regression. We can see that increasing heat has a small but highly significant effect on the overall shot difficulty. In particular, "hotter" players tend to take more difficult shots. Converting one more of the past 4 shots is associated with a $\hat{P}$ value drop of 0.55 percentage points [31].

[29] Due to the large number of variables in the model, the training data is vulnerable to severe overfitting.

[30] The expected percentage can be calculated with aid of the shot difficulty model.

[31] Making one more of the past 4 shots means that the shooting percentage increases by 25 percentage points. Therefore, we have $25 * (-0.022) = -0.55$

Table 6: The Effect of Heat on Overall Shot Difficulty

	Dependent variable:
	$\hat{P}$
Heat	-0.022^{***}
	(0.002)
Constant	0.460^{***}
	(0.0005)
Observations	107,568
R^2	0.001
Adjusted R^2	0.001
Residual Std. Error	0.155 (df = 107566)
F Statistic	120.493^{***} (df = 1; 107566)
Note:	*p<0.1; **p<0.05; ***p<0.01

4.6 Testing for the Hot Hand

After confirming that shot difficulty indeed increases with heat, we now want to check for the existence of the hot hand once controlling for shot difficulty. First, I will consider a baseline model, where I check how the team's probability of making a shot varies with heat. I run the following regression model using OLS:

$$P(\text{Make}_{is}) = \alpha + \beta * (\text{Heat}_{is}) + \theta * (\text{Team Fixed Effects}_i). \tag{15}$$

This model is comparable to the work done by GVT (no control for shot difficulty) and therefore one would expect a similar result as in Section 4.2, i.e. a negative β.

After being equipped with a nice baseline, I can turn to a specification where I test for the shot difficulty-controlled hot hand:

$$P(Make_{is}) = \alpha + \beta * (Heat_{is}) + \gamma * \hat{P}_{is} \tag{16}$$

The team fixed-effects no longer need to be included in the model since $\hat{P}_{is}$ already encapsulates them.

If the hot hand on a team level would not exist, we would expect the shot difficulty to be the only predictor of $P(Make)$. Hence, we would expect $\alpha = 0$, $\beta = 0$, and $\gamma = 1$. If the hot hand

does exist, however, we would expect a β significantly greater than zero.
Table 7 shows the results for both regressions. Notice that in Column 1 the coefficients for the *team fixed effects* were omitted due to the large number of coefficients and the fact that the main focus lies on the sign of β. Consistent with the observations made in Section 4.2, our baseline model reports a negative β (see Column 1). Upon controlling for shot difficulty, as expected, the β switches signs (see Column 2). Although β is significant, its effect size is very modest. To quantify the effect, if a team makes one more of its last four shots, the shooting percentage increases by 0.025 percentage points. Given that the average NBA team hits 46% of its shots (see Table 5) this corresponds to an 0.05 % improvement.

Table 7: The Hot Hand on the Team Level: Controlled for Shot Difficulty

	Dependent variable:	
	P(Make)	
	(1)	(2)
Heat	-0.019^{***}	0.001^{***}
	(0.002)	(0.0002)
$\hat{P}$		0.989^{***}
		(0.0003)
Constant	0.443^{***}	0.005^{***}
	(0.003)	(0.0001)
Observations	107,568	107,568
R^2	0.010	0.990
Adjusted R^2	0.010	0.990
Residual Std. Error	0.154 (df = 107537)	0.015 (df = 107565)
F Statistic	35.997^{***} (df = 30; 107537)	$5,485,503.000^{***}$ (df = 2; 107565)
Note:		*p<0.1; **p<0.05; ***p<0.01

4.7 Discussion

This subsection is dedicated to discussing several drawbacks and potential problems about the above used model, as well as possible modifications to improve its statistical power.
As already mentioned above, the most significant drawback is the incompleteness of the shot probability model. The absence of defensive controls, arguably the most influential factors on shot difficulty, might heavily affect the explanatory power of $\hat{P}$. Every basketball player knows that generally, the tighter one is guarded and the taller the defender, the more difficult the shot. Furthermore, next to team fixed effects, it certainly also makes sense to control

for differences between players[32]. If two-time MVP Stephen Curry and Deandre Jordan take identical jump shots, they have different probabilities to go in[33]. Moreover, the discussed model does not account for certain preferences of players in certain types of shots. James Harden, for example, is infamous for his lethal step-back jump shot. Since he specializes in this type of shot, the predicted shot probability for such a shot should be higher for him compared to the rest of the league. Similarly, it is also concerning that team fixed effects might not be precise enough to estimate $\hat{P}$ for individual teams. They do not account for the fact that certain teams specialize in certain plays and shots. For instance, under Coach Mike D'Antoni, the Houston Rockets became a team which tried to limit their shot selection to threes and layups. Finally, one could also question the use of an OLS regression. Although all coefficients in the discussed regressions are highly significant, they are nevertheless small in size (see Tables 6 and 7). Therefore, it is more likely that the estimated standard errors are inaccurate than the mean effects estimates. A more accurate model might alter the results (in either direction). It is plausible that incorporating the above mentioned changes reveal a more powerful effect of heat on the shooting probability. On the other hand, the possibility that any effect of heat vanishes altogether, cannot be ruled out either.

Nonetheless, the purpose of this section was also to spark a discussion about the possible existence of a hot hand on the team level. If it truly exists on a team level, it would be rational for teams/coaches to adjust their game strategy accordingly. For example, finding the right time to call a timeout to break a potential hot hand of the opposing team would become even more crucial than it already is. Furthermore, coaches might try to switch the defense of the whole team (rather than an individual player) in order to force the opposing team to adjust their offense[34]. In any case, it is certain that a quantifiable degree of hot hand shooting on a team level (no matter the extent) would be a valuable asset for coaches when making in-game adjustments and decisions.

5 Conclusion

For the better part of three decades, the hot hand was thought to be a myth in the majority of the scientific community. Recent research by Miller and Sanjurjo (2014, 2016, 2018b), however, called this belief into question as they point out a flaw in the empirical approach of Gilovich et al. (1985), the seminal paper on the hot hand fallacy. Said flaw is a selection bias in the estimator GVT used for their analysis. The first part of this thesis provided a summary of the

[32] Notice that the shot difficulty model has nothing to do with the definition of heat on a team level. It only tries to predict the conversion probability of a specific shot.

[33] Stephen Curry is by many regarded as the greatest (distance) shooter the sport of basketball has ever seen. Deandre Jordan is known for his spectacular dunks and ability to rebound. Therefore, the strengths these two players possess could not be any more different.

[34] In basketball, various defensive strategies and formations are used by pro teams. On a most general level one can differentiate between zone defense and man-to-man defense. Switching between those types of defenses forces the opposing team to restructure their offense and use a completely different playbook. This in turn could lead to an interruption of offensive "flow" in the opposing team.

nature of the bias and provided a sense of intuition on how GVT's approach is vulnerable to the bias. In the middle part, I extensively discussed my implementation of algorithms in R, which make it possible to quantify the bias in sequential binary data. We saw that the size of the bias remains quite substantial in common empirical frameworks. Upon correcting for the bias (using these algorithms), one can see that the results of GVT actually point towards the existence of hot hand shooting in basketball. The final part of this thesis introduced an approach to examining the hot hand on a team level. I find a small yet significant effect of hot hand shooting on a team level that might have noteworthy implications for in-game strategy adjustments of basketball teams.

Appendix G (Graphs)

This appendix lists graphs referenced in the thesis.

Figure 5: `Comparison of the Bias in` $\mathbb{E}[\hat{P}_k(\mathtt{X}) \mid I_k(\mathtt{X}) \neq \emptyset]$ `and` $\mathbb{E}[1 - \hat{Q}_k(\mathtt{X}) \mid J_k(\mathtt{X}) \neq \emptyset]$

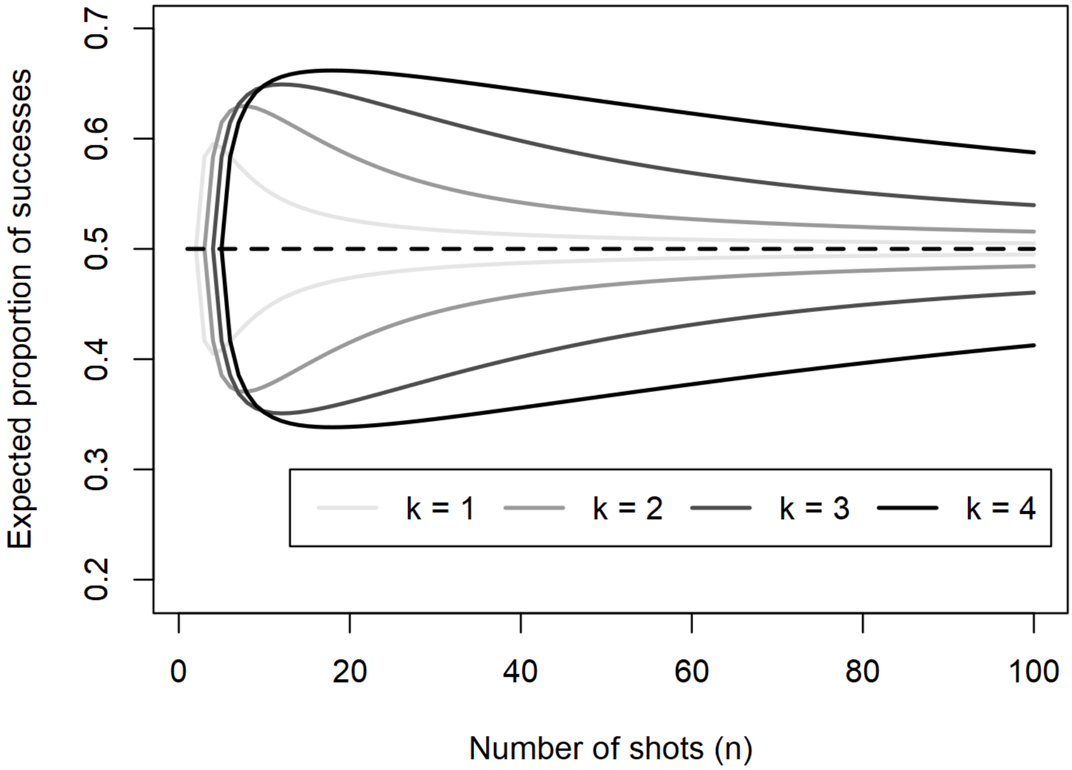


Figure 6: Bias-Corrected Difference for k = 2

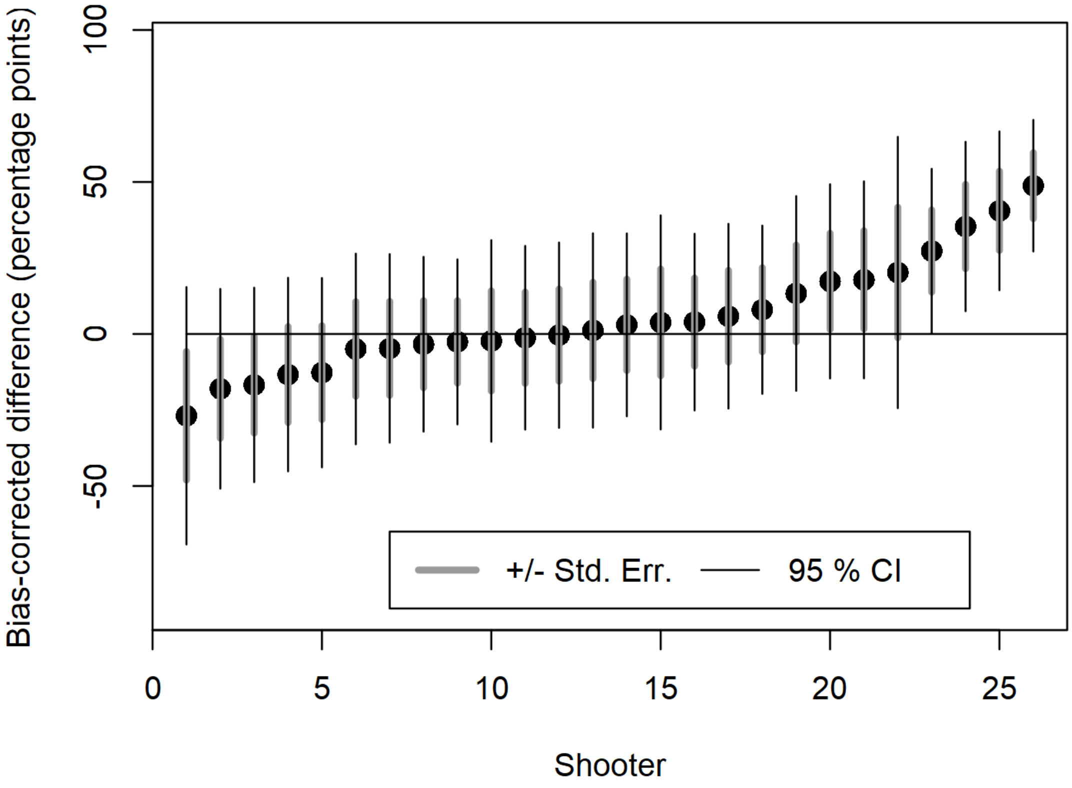
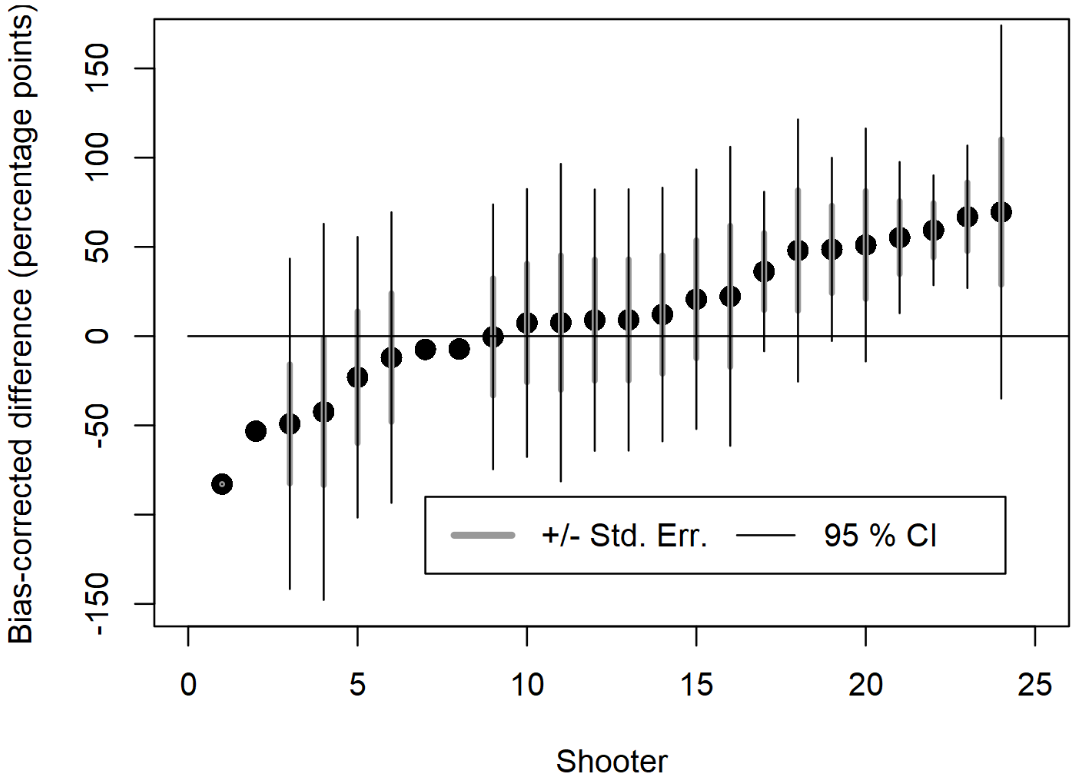


Figure 7: `Bias-Corrected Difference for k = 4`



Appendix R (R code)

This appendix will deal with the implementation of various algorithms in R, which were discussed throughout the thesis.

Algorithm 1 (`Helper Function 1`)
This function implements the operation $D^{\mathbf{m}':p'}$ for a dictionary D, a count realization $\mathbf{m}'$, and a probability p'.

```
update_power <- function(dict, m_prime, p_prime) {
    #updating dictionary
    for (i in 1:length(dict$keys)) {
        #add m_prime to each key
        dict$keys[[i]] <- dict$keys[[i]] + m_prime
    }
    b <- rep(p_prime, length(dict$keys))
    #scaling the probabilities
    dict$values <- dict$values * b
    return(dict)
    }
```

Algorithm 2 (`Helper Function 2`)
This function implements the operation $A \uplus B$ for two dictionaries A and B.

```
update_plus <- function(dict1, dict2) {
  # creating and returning the combined dictionary
  # Case 1: the dictionaries intersection is empty
  if (length(intersect(dict1$keys, dict2$keys)) == 0) {
    dict3 <- list(keys = c(dict1$keys, dict2$keys),
                  values = c(dict1$values, dict2$values))
    return(dict3)
  }
  # Case 2: the dictionaries have common keys
  else {
    c <- intersect(dict1$keys, dict2$keys)
    d <- match(c, dict1$keys)
    e <- match(c, dict2$keys)
    val <- dict1$values[d] + dict2$values[e]
    dict3 <- list(keys = c(c, dict1$keys[-d], dict2$keys[-e]),
                  values = c(val, dict1$values[-d], dict2$values[-e]))
    return(dict3)
  }
}
```

On a technical note, the programming language R does not support a *dictionary* data structure, at least not one that is equivalent to the one that languages like Python or C++ carry. To overcome this problem I have treated a named list, consisting of a list of keys and a vector of values/probabilities just like a dictionary as it is described in Section 3.1. The list of keys consists of vectors with two entries (failures, successes), which comes in very handy in Algorithm 1 because vector addition is needed there.

Algorithm 3 (`Dictionaries`)
This algorithm implements Algorithm 1 of Miller and Sanjurjo (2018a) in R. It builds a collection of dictionaries D. Of interest are especially the dictionaries D(0,n) for $n = k + 1, \ldots, N$, as these correspond to the joint distribution of the total number of (failures, successes) that immediately follow k consecutive successes in n trials.

```r
Count_Distribution <- function(N, k, p) {
  D <- as.list(numeric((k + 1) * (N + 1)))
  dim(D) <- c(k + 1, N + 1)
  q <- 1 - p
  for (n in 0:N) {
    L <- min(n, k)
    for (l in L:0) {
      r <- n - l
      if (r == 0) {
        D[[l + 1, r + 1]] <- list(keys = list(c(0, 0)), values = 1)
      }
      else if (r > 0) {
        if (l < k) {
          D[[l + 1, r + 1]] <- update_plus(update_power(D[[1, r]], c(0, 0), q),
                                           update_power(D[[l + 2, r]], c(0, 0), p))
        }
        else if (l == k) {
          D[[l + 1, r + 1]] <- update_plus(update_power(D[[1, r]], c(1, 0), q),
                                           update_power(D[[k + 1, r]], c(0, 1), p))
        }
      }
    }
  }
  return(D)
}
```

As one can see, the dictionaries are stored in a matrix. As opposed to python or C++, where indexing starts at 0, in R indexing starts at 1. This means, when you want to access the element in the first row and first column of a matrix M in R, then you execute M[1, 1] (as opposed to M[0,0] in python). For our situation at hand this means that the dictionary $D(0, n)$ cannot be stored in a place with the indices 0 and n. The natural solution is to shift the indices by one. So the dictionary $D(i, j)$ will just be stored in $D(i + 1, j + 1)$ in R. This also means that when working in R we are actually interested in the dictionaries $D(1, n + 1)$ for $n = k + 1, \ldots, N$. We have to keep this in mind when looking at the next algorithm, which calculates the exact expected proportion of successes on trials that were immediately preceded by k successes.

Algorithm 4 (`Expected Proportion of Successes 1`)
This function implements expected proportion of successes on trials that were immediately preceded by k successive successes.

```
exp_prop <- function(N, k, p) {
  # find out which sample of outcomes is relevant
  D <- Count_Distribution(N, k, p)
  # find relevant dictionary
  D_rel <- D[[1, N + 1]]
  # finding list element wich has key (0,0)
  for (i in 1:length(D_rel$keys)) {
    if (D_rel$keys[[i]][1] == 0 && D_rel$keys[[i]][2] == 0) {
      a <- i #storing the index of the list element with key (0,0)
    }
    else {
      next
    }
  }
  # calculating the denominator
  den <- 1 - D_rel$values[a]
  # calculating the numerator
  num <- 0
  for (i in 1:length(D_rel$keys)) {
    # skipping counts with zero successes as the coefficient is
    # zero in that case
    if (D_rel$keys[[i]][2] == 0) {
      next
    }
    else {
      # calculating the coefficient
      coeff <- D_rel$keys[[i]][2]/(D_rel$keys[[i]][1] + D_rel$keys[[i]][2])
      # multiplying the coefficient by the corresponding probability and
      # summing up
      num <- num + (coeff * D_rel$values[[i]])
    }
  }
  return(num / den)
}
```

Notice that the denominator is defined as $1 - p_D((0,0))$, which is of course equal to $\sum_{\mathbf{m}' \in D_c^*} p_D(\mathbf{m}')$.

Algorithm 5 `Expected Proportion of Successes 2`
This function implements the expected proportion of successes on trials that were immediately preceded by k successive failures.

```
exp_prop2 <- function(N, k, p) {
    return(1 - exp_prop(N, k, 1 - p))
}
```

Due to symmetry, in order to calculate exp_prop2 we can just calculate exp_prop2 but with the failure probability ($q = 1 - p$) as the input for p. Then we just subtract that result from one to obtain the expected proportion of successes following k consecutive failures.

Algorithm 6 (Expected Difference in Proportions)
This function gives back the expected difference in the probability of success when comparing trials that immediately follow k consecutive successes with trials that immediately follow k consecutive failures.

```
exp_diff <- function(N, k, p) {
  # find out which sample of outcomes is relevant
  D <- Count_Distribution_diff(N, k, p)
  # find relevant dictionary
  D_rel <- D[[1, 1, N + 1]]
  # finding list Element which has key (0,0)
  a <- c()
  for (i in 1:length(D_rel$keys)) {
    if ((D_rel$keys[[i]][1] == 0 && D_rel$keys[[i]][2] == 0) ||
        (D_rel$keys[[i]][3] == 0 && D_rel$keys[[i]][4] == 0)) {
      a <- c(a, i)
    }
    else {
      next
    }
  }
  # calculating the denominator
  den <- 1 - sum(D_rel$values[a])
  num1 <- 0
  for (i in 1:length(D_rel$keys)) {
    if ((D_rel$keys[[i]][2] == 0) ||
        (D_rel$keys[[i]][3] == 0 && D_rel$keys[[i]][4] == 0)) {
      next
    }
    else {
      coeff1 <- D_rel$keys[[i]][2] / (D_rel$keys[[i]][1] + D_rel$keys[[i]][2])
      num1 <- num1 + (coeff1 * D_rel$values[[i]])
    }
  }
  num2 <- 0
  for (i in 1:length(D_rel$keys)) {
    if ((D_rel$keys[[i]][4] == 0) ||
        (D_rel$keys[[i]][1] == 0 && D_rel$keys[[i]][2] == 0)) {
      next
    }
    else {
      coeff2 <- D_rel$keys[[i]][4] / (D_rel$keys[[i]][3] + D_rel$keys[[i]][4])
      num2 <- num2 + (coeff2 * D_rel$values[[i]])
    }
  }
  num <- num2 - num1
  return(num / den)
}
```

Appendix S (Statistical Background)

This appendix provides some statistical background for the calculations made in Section 3.2. For an estimator $\hat{\beta}$ of a parameter β the t-statistic for this parameter is calculated through

$$t(\hat{\beta}) = \frac{\hat{\beta} - \beta_0}{SE(\hat{\beta})},$$

where SE stands for the standard error, and β_0 is the value β takes under the null hypothesis $(H_0 : \beta = \beta_0)$. For our case at hand $(k = 3)$[35], we have $\beta_0 = 0$, β is the bias-adjusted difference in proportions, and $\hat{\beta}$ is the mean of the bias-adjusted difference in proportions (0.126). $SE(\hat{\beta})$ is therefore just the standard error of the mean, which can be calculated through

$$SE(\hat{\beta}) = \frac{\sigma}{\sqrt{n}},$$

where σ is the standard deviation in the sample of differences and n is the sample size. The standard deviation of Column 7 of Table 4 is now computed in a straightforward manner. In R, for example, with the builtin function sd(). Here comes the drawback mentioned in Section 3.2 into play. Notice that every adjusted estimate in the calculation of the standard deviation possesses an equal weight and ignores the number of shots each player took in the respective categories ''3 makes" and ''3 misses". The standard deviation is then used to calculate the standard error of the mean, which is then used to calculate the T-statistic and the p-value. This is the reason why Miller and Sanjurjo (2018b) argue that this t-test limits statistical power and propose an alternative way of calculating the standard error and p-value of the test statistic. Under the assumption of normality and for a one-sided (right-tailed) test the p-value can be calculated as

$$p = 1 - \text{pnorm}\left(\frac{\hat{\beta} - \beta_0}{SE(\hat{\beta})}\right),$$

where *pnorm* is the probability density function of the standard normal distribution and $\hat{\beta}$, β, and β_0 defined as above.

[35] The other cases $(k = 2, 4)$ work of course analogously.

Appendix T (Tables)

This appendix provides tables referenced in the thesis.

Table 8: Shooting Performance and Adjusting the Bias for $k = 2$

shooter	$\hat{P}$(hit|2 makes)	$\hat{P}$(hit)	$\hat{P}$(hit|2 misses)	GVT est. ($\hat{D}_2$)	bias$_2$	bias adj. ($\hat{A}_2$)
Males						
1	0.48 (25)	0.54 (100)	0.5 (18)	-0.020	-0.032	0.012
2	0.25 (12)	0.35 (100)	0.333 (42)	-0.083	-0.036	-0.047
3	0.625 (40)	0.6 (100)	0.684 (19)	-0.059	-0.033	-0.026
4	0.231 (13)	0.4 (90)	0.448 (29)	-0.218	-0.037	-0.180
5	0.4 (15)	0.42 (100)	0.6 (30)	-0.200	-0.033	-0.167
6	0.622 (37)	0.57 (100)	0.381 (21)	0.241	-0.032	0.273
7	0.63 (27)	0.56 (75)	0.5 (16)	0.130	-0.043	0.173
8	0.636 (11)	0.5 (50)	0.5 (12)	0.136	-0.066	0.202
9	0.789 (38)	0.54 (100)	0.333 (30)	0.456	-0.032	0.488
10	0.6 (35)	0.6 (100)	0.5 (14)	0.100	-0.033	0.133
11	0.618 (34)	0.58 (100)	0.611 (18)	0.007	-0.033	0.039
12	0.389 (18)	0.44 (100)	0.433 (30)	-0.044	-0.032	-0.012
13	0.559 (34)	0.61 (100)	0.615 (13)	-0.057	-0.034	-0.023
14	0.588 (34)	0.59 (100)	0.625 (16)	-0.037	-0.033	-0.004
Females						
1	0.45 (20)	0.48 (100)	0.609 (23)	-0.159	-0.031	-0.127
2	0.357 (14)	0.34 (100)	0.364 (44)	-0.006	-0.037	0.030
3	0.421 (19)	0.39 (100)	0.375 (40)	0.046	-0.034	0.080
4	0.333 (9)	0.32 (100)	0.333 (45)	0.000	-0.038	0.038
5	0.5 (12)	0.36 (100)	0.357 (42)	0.143	-0.035	0.178
6	0.412 (17)	0.46 (100)	0.577 (26)	-0.165	-0.032	-0.133
7	0.65 (20)	0.41 (100)	0.278 (36)	0.372	-0.033	0.405
8	0.577 (26)	0.53 (100)	0.55 (20)	0.027	-0.032	0.058
9	0.471 (17)	0.45 (100)	0.552 (29)	-0.081	-0.032	-0.049
10	0.667 (21)	0.46 (100)	0.345 (29)	0.322	-0.032	0.354
11	0.5 (28)	0.53 (100)	0.565 (23)	-0.065	-0.032	-0.034
12	0 (5)	0.25 (100)	0.315 (54)	-0.315	-0.046	-0.269
Average	0.491	0.472	0.472	0.018	-.035	.053

Table 9: `Shooting Performance and Adjusting the Bias for` $k = 4$

shooter	$\hat{P}(\text{hit}\|4\text{ makes})$	$\hat{P}(\text{hit})$	$\hat{P}(\text{hit}\|4\text{ misses})$	GVT est. ($\hat{D}_4$)	bias_4	bias adj. ($\hat{A}_4$)
1	0.5 (6)	0.54 (100)	0.6 (5)	-0.100	-0.176	0.076
2	- (0)	0.35 (100)	0.375 (16)			
3	0.6 (15)	0.6 (100)	1 (2)	-0.400	-0.172	-0.228
4	0 (1)	0.4 (90)	0.714 (7)	-0.714	-0.183	-0.531
5	0 (2)	0.42 (100)	1 (3)	-1.000	-0.174	-0.826
6	0.6 (15)	0.57 (100)	0.222 (9)	0.378	-0.174	0.552
7	0.545 (11)	0.56 (75)	0.25 (4)	0.295	-0.215	0.511
8	0.75 (4)	0.5 (50)	0.333 (3)	0.417	-0.280	0.696
9	0.8 (25)	0.54 (100)	0.385 (13)	0.415	-0.176	0.592
10	0.583 (12)	0.6 (100)	0.667 (3)	-0.083	-0.172	0.088
11	0.615 (13)	0.58 (100)	0.667 (3)	-0.051	-0.174	0.122
12	0.333 (3)	0.44 (100)	0.3 (10)	0.033	-0.175	0.209
13	0.375 (8)	0.61 (100)	0.667 (3)	-0.292	-0.170	-0.121
14	0.583 (12)	0.59 (100)	0.667 (3)	-0.083	-0.173	0.089
1	0 (3)	0.48 (100)	0.667 (3)	-0.667	-0.177	-0.490
2	0.5 (2)	0.34 (100)	0.438 (16)	0.062	-0.161	0.223
3	0.5 (4)	0.39 (100)	0.188 (16)	0.312	-0.170	0.483
4	0 (1)	0.32 (100)	0.227 (22)	-0.227	-0.156	-0.071
5	0 (1)	0.36 (100)	0.238 (21)	-0.238	-0.165	-0.073
6	0 (2)	0.46 (100)	0.6 (5)	-0.600	-0.176	-0.424
7	0.5 (8)	0.41 (100)	0.312 (16)	0.188	-0.173	0.360
8	0.636 (11)	0.53 (100)	0.333 (3)	0.303	-0.176	0.480
9	0.25 (4)	0.45 (100)	0.429 (7)	-0.179	-0.176	-0.003
10	0.8 (10)	0.46 (100)	0.308 (13)	0.492	-0.176	0.668
11	0.5 (4)	0.53 (100)	0.6 (5)	-0.100	-0.176	0.076
12	- (0)	0.25 (100)	0.28 (25)			
Average	0.416	0.472	0.479	-0.077	-0.179	0.102

Table 10: `Variables Used in the Shot Prediction Model`

name	type	description
`shot_distance`	numeric	Shot distance in feet
`shot_distance2`	numeric	Shot distance squared
`shot_distance3`	numeric	Shot distance cubed
`play_length`	numeric	Length of the possession where the shot was recorded in seconds
`remaining_time`	numeric	Remaining time in the quarter (in seconds) at the time of the shot
`score_diff`	numeric	Absolute value of the score differential between the teams at the time of the shot
`fastbreak`	binary	Indicates if the shot came out of a fastbreak situation (`play_length` was less or equal to 7 seconds) or not.
`team`	categorical	Indicates which of the 30 NBA teams took the shot
`period`	categorical	Period in which the shot occurred (1-5). Overtime was regarded as a fifth period
`forced`	binary	Was the shot was taken within the last two seconds of a quarter.
`shot_type`	categorical	The type of shot. The levels are: Dunk, Layup, 3 Point-Shot, Hook Shot and Jump Shot.
`shot_adddiff`	categorical	Added difficulty of the shot, i.e. Fadeaway, Putback, Step-Back, Reverse etc. A total of 15 levels are considered
`home`	binary	Indicates if the team who shot was playing at home ($=$ 1) or away ($=$ 0)

References

Bocskocsky, A., J. Ezekowitz, and C. Stein (2014): "The Hot Hand: A New Approach to an Old 'Fallacy'," in *8th Annual MIT Sloan Sports Analytics Conference*, 1–10.

Carlson, K. A. and S. B. Shu (2007): "The Rule of Three: How the Third Event Signals the Emergence of a Streak," *Organizational Behavior and Human Decision Processes*, 104, 113–121.

Fisher, Sam (2015): "The Hot Hand Fallacy Is Safe," `http://www.rightcallconsulting.com/the-hot-hand-fallacy-is-safe/`, [Online; accessed 27-June-2019].

Gelman, Andrew (2016): "Miller and Sanjurjo Share 5 Tips on How to Hit the Zeitgeist Jackpot," `https://statmodeling.stat.columbia.edu/2016/02/18/miller-and-sanjurjo-share-5-tips-on-how-to-hit-the-zeitgeist-jackpot/`, [Online; accessed 27-June-2019].

Gilovich, T., R. Vallone, and A. Tversky (1985): "The Hot Hand in Basketball: On the Misperception of Random Sequences," *Cognitive psychology*, 17, 295–314.

Miller, J. B. and A. Sanjurjo (2014): "A Cold Shower for the Hot Hand Fallacy," Working paper, availalable at https://dx.doi.org/10.2139/ssrn.2450479.

——— (2016): "A Primer and Frequently Asked Questions for 'Surprised by the Gamblers and Hot Hand Fallacies? A Truth in the Law of Small Numbers' (Miller and Sanjurjo 2015)," Online Resource, availalable at https://dx.doi.org/10.2139/ssrn.2728151.

——— (2017): "A Bridge from Monty Hall to the Hot Hand: Restricted Choice, Selection Bias, and Empirical Practice," .

——— (2018a): "Supplement to 'Surprised by the Hot Hand Fallacy? A Truth in the Law of Small Numbers'," *Econometrica Supplemental Material*, 86, 2019–2047.

——— (2018b): "Surprised by the Hot Hand Fallacy? A Truth in the Law of Small Numbers," *Econometrica*, 86, 2019–2047.

NBA Media Ventures, LLC (2019): "NBA Statistics," https://stats.nba.com/player/202691/?Season=2018-19&SeasonType=Regular%20Season, [Online; accessed 23-May-2019].

Rao, J. M. (2009): "Experts' Perceptions of Autocorrelation: The Hot Hand Fallacy Among Professional Basketball Players," Working paper, availalable at https://pdfs.semanticscholar.org/2f15/0b534bfa31a2c439eb9b0ebc2051faae8d9b.pdf.

Schäfer, N. (2019): "Hot-Hand," GitHub Repository, available at https://github.com/nschaefer1211/Hot-Hand.

Stone, D. F. (2012): "Measurement Error and the Hot Hand," *The American Statistician*, 66, 61–66.

Wardrop, R. L. (1999): "Statistical Tests for the Hot-Hand in Basketball in a Controlled Setting," *American Statistician*, 1, 1–20.