MATH 1610 — Fall 2015 **Final Project** December 19, 2015

Analysis of 3 Point Shooting in the NBA for the 2013-2014 Season

Blake Mandell

December 6, 2015

Abstract

This project investigated the modeling of two data points that describe three point shooting: the number of three point field goal attempts attempted and the percentage with which a player made his three point attempts. The first undertaking dealt with fitting different distributions to three-point percentages and three-point attempts, and then creating a measure of how well each different distribution fits the data. The second undertaking attempts to approximate how close a player's 3-point shot attempts and 3-point shot makes are to an independent random variable through the use of modeling data toward a Poisson distribution.

While there were 481 players who played in the NBA during the 2013-2014 season, only 106 of them – the qualified players – made 82 three-point field goals (an average of one per game in an 82 game season). This project focused exclusively on the three-pointers attempted and the three-point percentages for the qualified players.

Each of the two aforementioned data points were graphed as histograms and then fitted and compared to various probability distributions: the normal (Gaussian) distribution, the Beta distribution, the Cauchy distribution, and the Laplace distribution. Pearson product-moment correlation coefficient ($r = 0.48533$) was also found for the number of three-point attempts and the three-point percentages of qualified players.

Is shooting an independent random variable? More specifically, how well can the number of three-point shots that a player makes and the number of three-point shots that a player attempts modelled by a Bernoulli trials process each time that the player comes down the floor? I utilize the Poisson distribution and how well a player's shooting data matches the fitted distribution to begin to formalize how far away a player's shooting is from being modelled by an independent random variable from possession to possession.

While the process can be automated for more players, the second undertaking was more mathematically involved than the first. Thus, I only attempted to mathematically

elucidate the notion of a player's independence in shooting for four different players who are all high-volume shooters or three-point shooting specialists: Kobe Bryant, Steph Curry, Kyle Korver, and James Harden.

Method I

As a precursor, all data – the total statistics for each player in the 2013-2014 NBA season – for this project was downloaded from Basketball Reference, where it was downloaded as a CSV file after manually removing each row that had its first column's value as "Rk" instead of a unique player's number. All code was written in Julia, and will be sent along with this paper.

I then parsed through the downloaded CSV file to output all qualified players' (as mentioned before, a player must have made 82 three-point shots in a season to have qualified three-point statistics) three-point attempts and percentages.

In another file, I began to do the main work of this project: to 1) attempt to model each of the two data sets by common probability distributions 2) see how well each distribution modeled the data 3) correlate each of the data sets. I shall go through each of these three objectives separately.

Before I modeled the data sets with probability distributions, I used Julia's built-in `hist(v[, n]) -> e, counts` function to take each data set and turn it into a histogram, where `e` represents the bins of the histogram and `counts` represents how many objects are in each bin. To graph in a slightly more aesthetic and appealing manner, I took the midpoints of each of the bins in `e` and used the Gadfly package to graph those histograms instead.

To graph each of the histograms relating to the number of three-point attempts and three-point percentages in relation to common probability distributions first required a scaling of each of the histograms. Initially, the histograms were graphed with the x values as the location of the bins and y values as how many players fit in each bin, but the total areas of those graphs didn't sum up to 1, which is necessary for a function to be a probability distribution. To scale each of the histograms correctly, I created a `stand_hist(all_data,bin_size)` that returned scaled data that would create a histogram with area 1.

After I scaled the histogram to have area 1, I used the Julia Distributions package for all distribution-related fitting and creation. More specifically, I used the `fit(distribution,data) -> fitted_distribution` function to fit the the normal (Gaussian) distribution, the Beta distribution, the Cauchy distribution, and the Laplace distribution to each of the scaled data. I then used the Gadfly library to graph each of these fitted distributions as a layer along with whatever scaled histogram (either the three-point attempts or three-point percentages) to which it referred. All the graphs are in the accompanying PDF file of all the code and outputs.

I used three measures to compare how well each fitted distribution modeled the scaled histogram data. I shall write each of them mathematically, and then describe them. Two tables – one for three-point attempts and the other for three-point percentages – will be given at the end of this write-up to display all of the data I created using distribution-fitting and my measures of distribution accuracy.

In my equations, J is the set of all bars in the histogram, $f(x_i)$ is the height of the histogram at any specific $i \in J$, and $p(x)$ is the height of a probability distribution at any point. For equation 3, R represents the argument `riemann` in the code and w represents the width of a histogram bin.

$$E_1 = \sum_j |f(x_j) - p(x_j)| \tag{1}$$

$$E_2 = \sum_j (f(x_j) - p(x_j))^2 \tag{2}$$

$$E_3 = cdf(p(x), min(J)) + (1 - cdf(p(x), max(J)) + \sum_{j \in J} \sum_{i=1}^{R} \int_{j+((i-1)/w)}^{j+(i/w)} pdf(x)dx \tag{3}$$

The first error equation represents the mean of the absolute values of the differences between the height of the scaled histogram at a certain point and the height of whichever fitted distribution I am using. For this equation, each variation from the used distribution is weighted equally.

The second error equation represents the mean of the square of the differences between the height of the scaled histogram at a certain point and the height of whichever fitted distribution I am using. For this equation, each variation from the used distribution is square, which weights larger variations significantly more than smaller variations. In other words, this error equation strongly weighs outliers.

The third equation attempts to resolve the inherent issue between a discrete histogram and a continuous probability distribution: namely, that the former is a discrete histogram that is "mapped" onto a continuous density function and then compared to a continuous probability distribution. To try to offset that, I use the idea of a Riemann sum to model the error. After getting the area under the fitted distribution before the first bar of the and after the last bar of the histogram, I measure the difference in the area between a small section of each histogram bar and the chosen distribution over the same x interval (more specifically, R times per bar of the histogram). Finally, I add everything together to find how much area the distributions don't share with each other as an error bound.

Finally, I computed the Pearson product-moment correlation coefficient, often called r, for three-point shots attempted and three-point shot percentages. It turns out that $r = 0.137299$, which means that just under 14 percent of three-point shots for qualified three-point shooters can be explained by linear relationship from three-point field goal percentage, and vice versa. The explanation for this is explained far more simply from a basketball understanding than in rigorous mathematics: for a player who has an average of making at least one three-point shot per game, it is not unlikely that he will make a three in a game. Thus, the defense will be aware that said player is an adequate three-point shooter, and will have a defensive aim to avoid letting that player roam the three-point line while open. Kirk Goldsberry terms this ”player gravity.” Furthermore, many of the best three-point shooters in the league in terms of percentage are essentially three-point specialists, meaning that they are not much of a scorer outside of the three-point shot. Thus, they are more of an auxiliary scorer than a primary one, and will not take as many three-point shots.

Tables I

3 Point Percentages and Fitted Distributions					
Distribution Type	Param 1	Param 2	E_1 Output	E_2 Output	E_3 Output
Normal	$\mu = 0.37435$	$\sigma = 0.03280$	76.7457	324.407	0.4020034
Beta	$\alpha = 80.3406$	$\beta = 134.2728$	77.003196	326.063229	0.40216
Cauchy	$\mu = 0.37437$	$\sigma = 0.020398$	80.6136	401.87644	0.563234
Laplace	$\mu = 0.37437$	$\theta = 0.028604$	74.316291	361.0576	0.454228

3 Point Attempts and Fitted Distributions					
Distribution Type	Param 1	Param 2	E_1 Output	E_2 Output	E_3 Output
Normal	$\mu = 355.2075$	$\sigma = 106.7441$	0.0617016	0.0001666	0.678909
Cauchy	$\mu = 328.0$	$\sigma = 72.625$	0.058041	0.0001671	0.813935
Laplace	$\mu = 328.0$	$\theta = 105.2646$	0.058019	0.0001597	0.742178

Method II

Because this undertaking involves all three of colloquial basketball knowledge regarding ”streakiness” and an accompanying folk theory, probability theory regarding the linkage between the binomial and Poisson distributions, and actual programming, I will work through all three, one by one.

The folk theory of a streaky shooter in basketball is as follows: A player will attempt and make a few consecutive three-point shots, usually the beginning or end of a game. That player, now colloquially said to be ”hot” or ”heating up”, will continue to make shot after shot in a row because he is ”feeling it.” The opposite will also be true: if a player knows that he is a streaky shooter and misses a few shots in a row, he will continue missing more shots than usual for the duration of the game, as he is ”cold” during this game. The phenomena of in-game shooting streakiness has been researched in the past, particularly by Kochler and Conley and Csapo et. al.

A simple hypothetical story with real-world basketball players can illustrate the importance of verifying if the folk concept of streakiness actually exists mathematically: say that team A is down by 3 with 5 seconds left. A time out is called, and Team A has possession of the ball. Team A just so happens to have both JR Smith and Kyle Korver, who has one of the highest three-point shooting percentages of time at .435%. Just a few minutes ago, Team A was down by 15, but JR has made 4 three-point shots in a row. Numerically, JR Smith has a far lower 3-point shooting percentage than Kyle Korver. JR is, though, colloquially known to be the streakiest shooter in all of the NBA. If the folk concept of streakiness is true (meaning that in fact a player's shooting is not well-modelled as a Poisson random variable), then perhaps it would be wise for the coach to take into account the ”hot streak,” and plan a play for JR to shoot. If it's not (thus, that a player's shooting is well-modelled by a Poisson random variable), then the coach would be wise to create a play for one of the best three-point shooters of all time, Kyle Korver.

The impact of in-game streakiness as divergence from a player's three-point shooting and three-point attempts being modelled as an independent Poisson variable has not yet been studied. More explicitly, how well can a player's rate of three-point shots attempted and a player's rate of three-point shots made be modelled possession-by-possession by an independent random variable?

To provide a little mathematical background, a Bernoulli trial is a random experiment that has two outcomes, ”success” and ”failure.” ”Success” occurs with probability p, and failure occurs with probability $1-p$. In probability, a ”success” corresponds to the value 1, and ”failure” corresponds to the value 0.

The concept of independence is intuitively tied to into the idea of Bernoulli trials. If I am flipping a fair coin, it doesn't matter if it's my first flip or my 10th flip with only heads before this flip: previous coin flips do not affect the probability that this flip will be heads (50%) or tails (50%). More formally, two events A and B are independent if and only if their joint probabilities equals the product of their individual probabilities. More formally, $P(A \cap B) = P(A)P(B)$. More colloquially, the occurrence of A does not affect the probability of B, and the occurrence of B does not affect the probability of A.

What happens if I flip a coin of probability p multiple times, say n? How many heads will I have counted? It is trivial that I will have no fewer than 0 and no more than n. The binomial distribution informs me of the probability that I will see k heads in n

flips as

$$P(X=k)=\binom{n}{k}p^k(1-p)^{n-k}, \text{where}\binom{n}{k}=\frac{n!}{k!(n-k)!} \tag{4}$$

The binomial distribution requires arguments: n, the number of trials, and p, the probability that any given trial will be a success. As n increases to infinity and p remains relatively small, the binomial distribution converges to the Poisson distribution. The Poisson distribution, which needs only one input λ, expresses the probability of a given number of events occurring independently at a fixed average rate over an interval of time. For a given $\lambda > 0$, the Poisson distribution claims that there will k occurrences of the event with probability

$$P(X=k)=\frac{\lambda^k e^{-k}}{k!} \tag{5}$$

In fact, if n is large and p is small, then a binomial distribution can be approximated with a Poisson distribution of parameter $\lambda = n * p$. Let n be large, p be small, and $\lambda = np$:

$$\binom{n}{k}p^k(1-p)^{n-k}=\frac{n(n-1)...(n-k+1)}{k!}\left(\frac{\lambda}{n}\right)^k\left(1-\frac{\lambda}{n}\right)^{n-k} \tag{6}$$

$$\approx \frac{\lambda^k}{k!}\left(1-\frac{\lambda}{n}\right)^n \quad \text{if } k \text{ is small compared to } n \tag{7}$$

$$\approx \frac{\lambda^k}{k!}e^{-\lambda} \quad \text{if } n \text{ is large} \tag{8}$$

The average basketball game has somewhere around 100 possessions per team, and it is fairly uncommon for any one player to take more than 10 three-point attempts per game. Using these approximate measures, it seems that n is significantly larger than the percentage that any random takes a three-point shot on any possession down the floor: $P(\text{shot}) = \frac{\text{number of three-point shots attempted or made}}{\text{number of possessions}} < .1$. p will be small both for shots attempted and shots made. Thus, we can utilize the Poisson distribution with $\lambda = np$ to approximate the binomial distribution. This has the benefit of only having to fit three-point shot attempts and three-point shot makes to just one variable: λ.

What we are concerned about most in fitting data to a Poisson distribution is how much the data deviates from the distribution in question. Because the Poisson distribution depends upon the fact that the occurrences of a certain number of events depends entirely upon the events occurring independently (and with fixed λ), however much the data deviates from the Poisson distribution gives data about how the rate at which a player takes shots changes from game to game, and thus provides insight into how streaky a shooter is. The more that a data set deviates from its best-fit Poisson distribution, the less independent the data is. In other words, the more that the data deviates from its best-fit Poisson distribution, the less fixed is the rate with which a player shoots and makes three-point shots.

Before I can model a best-fit Poisson to the data, I need to process the data. This project attempts to formalize the notion of three-point shooting streakiness for four specific players: Kobe Bryant, Stephen Curry, Kyle Korver, and James Harden. All four of these players have been high-volume shooters or three-point specialists for the last few years in the league. I use Kobe's data from the last four complete seasons (2012-2015) due to games lost from injury, and the last three complete seasons (2013-2015) for the latter three players.

In processing the data, I first extract each player's number of three-point shots attempted and made per game, standardizing each datapoint by converting it to a per-36 minutes value. To do this, I divide the number of shots attempted or made by the number of minutes played that game, and then multiply that by 36. Each team plays at a different pace, defined as the number of possessions per 48 minutes, and all four of the players I used here as examples are on different teams during years I am investigating, so I need to convert possessions-per-48 to possessions-per-36. I use the average of each respective player's team's pace over as possessions-per-48. I then divide each per-36 minute shots taken data point by the number of possessions per 36 minutes to get an approximate number of shots per possession, and then multiply that by 100 to obtain an approximate number of three-point shots made and attempted per 100 possessions. Because I process each data point (representing one game) this way, I output a list of three-point shots made and attempted per 100 possessions.

Now, there are two lists of per-100-possessions data for each of the four selected players: one of three-point shots made, and one of three-point shots attempted. Similar to my process in Method II, I use Julia's `hist(v[, n]) -> e, counts` function to take each data set and turn it into a histogram, where `e` represents the bins of the histogram and `counts` represents how many objects are in each bin. I specify `n` as the `0:1:100`, which means every integer value between 0 and 100 inclusive. Thus, this histogram automatically deals with non-integer values and sorts them into the correct bin. To allow y values of `counts` to add up to 1, I divide `counts` by `sum(counts)` to scale the number of players per bin to something that can be modelled by a probability distribution.

Each of the scaled histogram data is now in a form that can be compared to a Poisson distribution. To find the best-fit Poisson distribution to each list of data, I utilize a function that finds the error between a Poisson distribution with fixed λ and the data set – thus, through trial and error, I find the best-fit λ by minimizing the error E_P (error for Poisson), which is defined with data v and Poisson parameter λ as follows:

$$E_P\left(v,\lambda\right) = \textstyle\sum_{k=0}^{\text{length}(v)} \text{abs}\left(v(k) - \frac{e^{-\lambda}\lambda^k}{k!}\right) \tag{9}$$

Once a best-fit Poisson distribution is found for each of the two data types (three-point shots attempted and three-point shots made) for each of the four players, a final E_P

is computed to measure each data's deviance from the best-fit Poisson distribution. Even if a player's three-point shots attempted perfectly matches a Poisson distribution, he can still be a streaky shooter if his three-points shots made deviate from their own best-fit Poisson distribution. What is important to the notion of "streakiness" (at least the relatively naive notion developed in this paper) is how much more a player's data deviates from the best-fit shots-made Poisson distribution than the best-fit shots-attempted Poisson distribution. Thus, for this paper, a rough mathematical approximation of "streakiness" will be defined as follows:

$$\text{Streakiness} = E_P(v_{\text{made shots}}, \lambda_{\text{made shots}}) - E_P(v_{\text{attempted shots}}, \lambda_{\text{attempted shots}}) \tag{10}$$

Unfortunately, there was an issue in the code with how the E_P function is used to create a best-fit Poisson distribution. While all of the made-shots data was able to be fit to a corresponding Poisson distribution with an intuition-matching λ value, only one of the player's data (Kobe) worked to create a reasonable λ value for a matching Poisson distribution. Thus, all four of the players' attempted shots best-fit Poisson distributions will be discarded, and no formal measure of streakiness will be reported. I will expound upon the errors in the Errors section, and will offer the four graphs that – to the eye test – look like Poisson distributions.

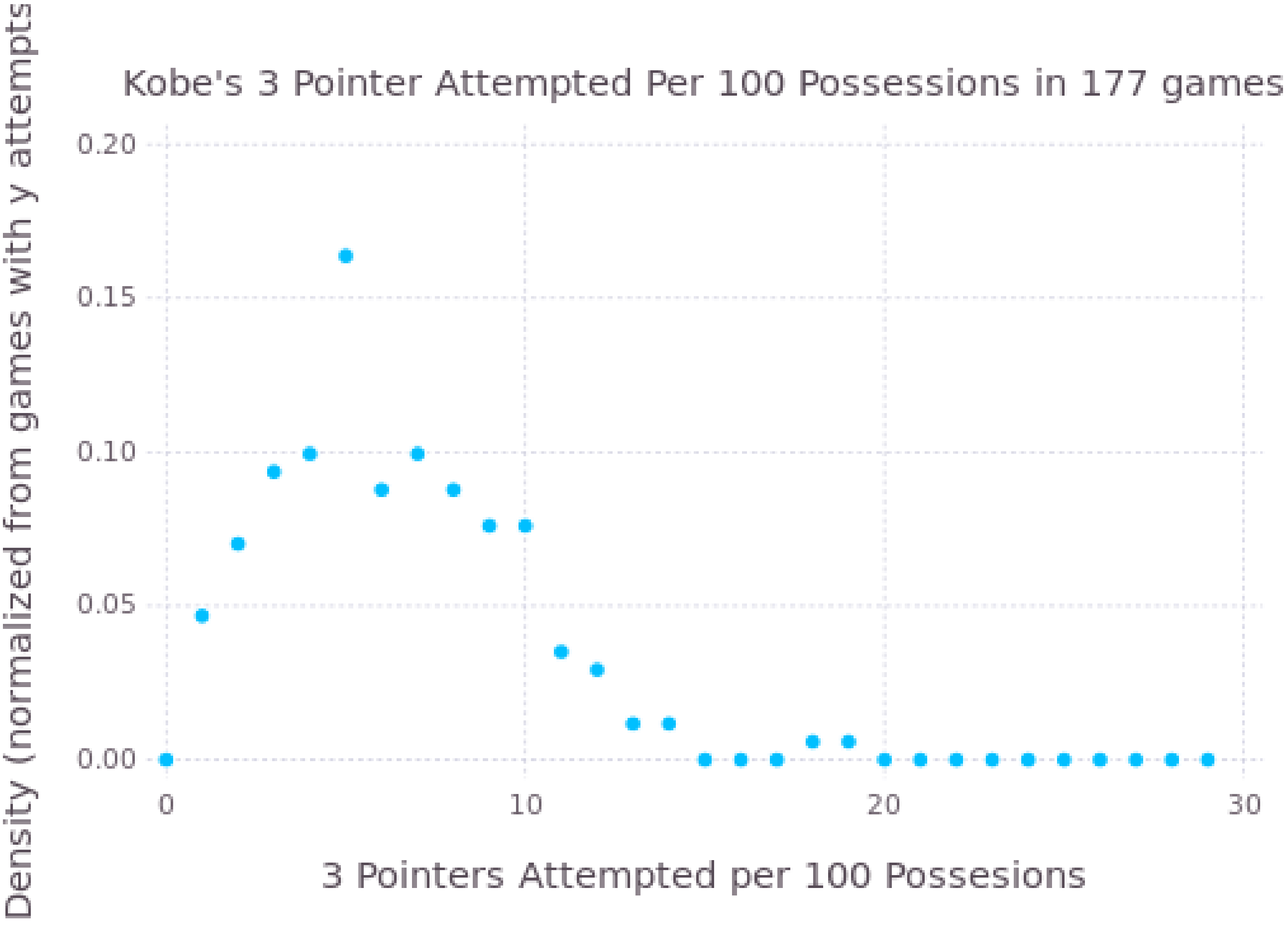
Fortunately, the E_P values that were computed for each of the player's made shots data matches the eye test. For the approximations and error bounds computed with made shots only, Kobe is the most streaky shooter ($E_P = 0.3454$) of the four, followed by Harden ($E_P = 0.3006$). Curry is the second least streaky ($E_P = 0.2789$), and Korver, as expected, is the least streaky shooter ($E_P = 0.2630$).

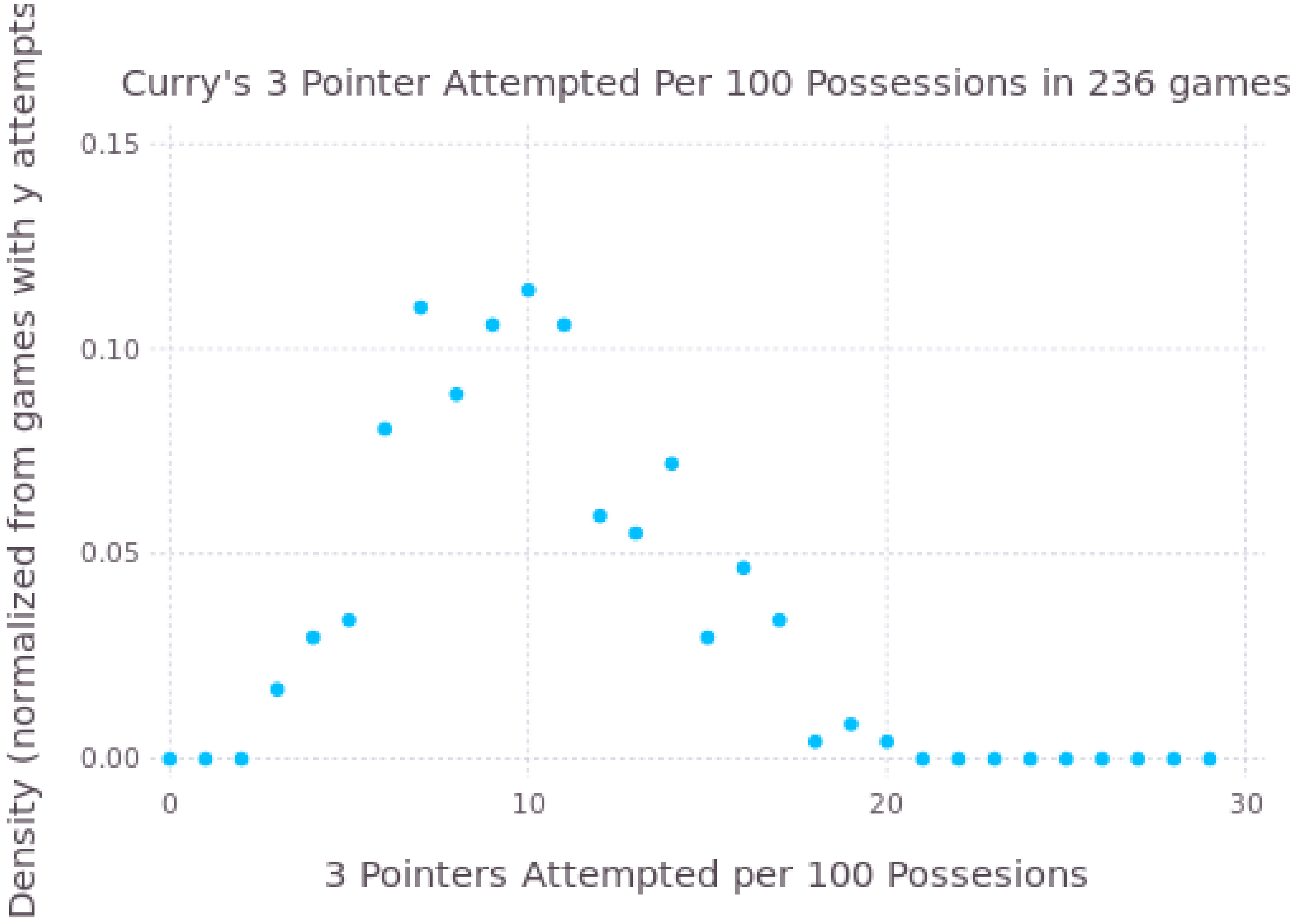
Table II

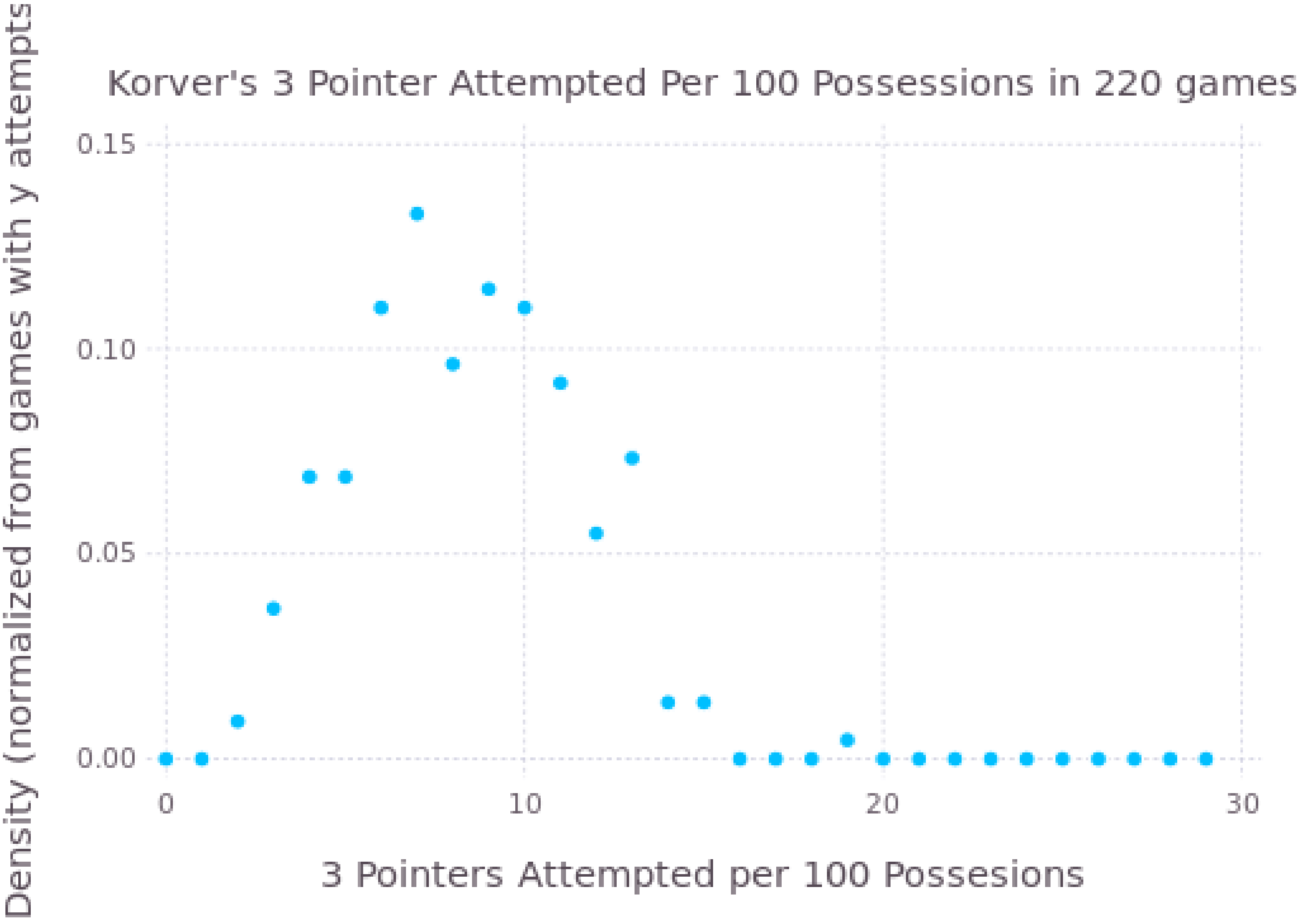
Poisson Distributions Fitted to Three-Point Shots Made and Attempted				
Player	λ_{made}	E_{made}	$\lambda_{\text{attempted}}$	$E_{\text{attempted}}$
Kobe	3.2227	0.3454	6.9414	0.2856
Curry	4.8574	0.2789	0.0	1.0
Korver	4.5137	0.2630	0.0	1.0
Harden	3.4590	0.3006	0.0	1.0

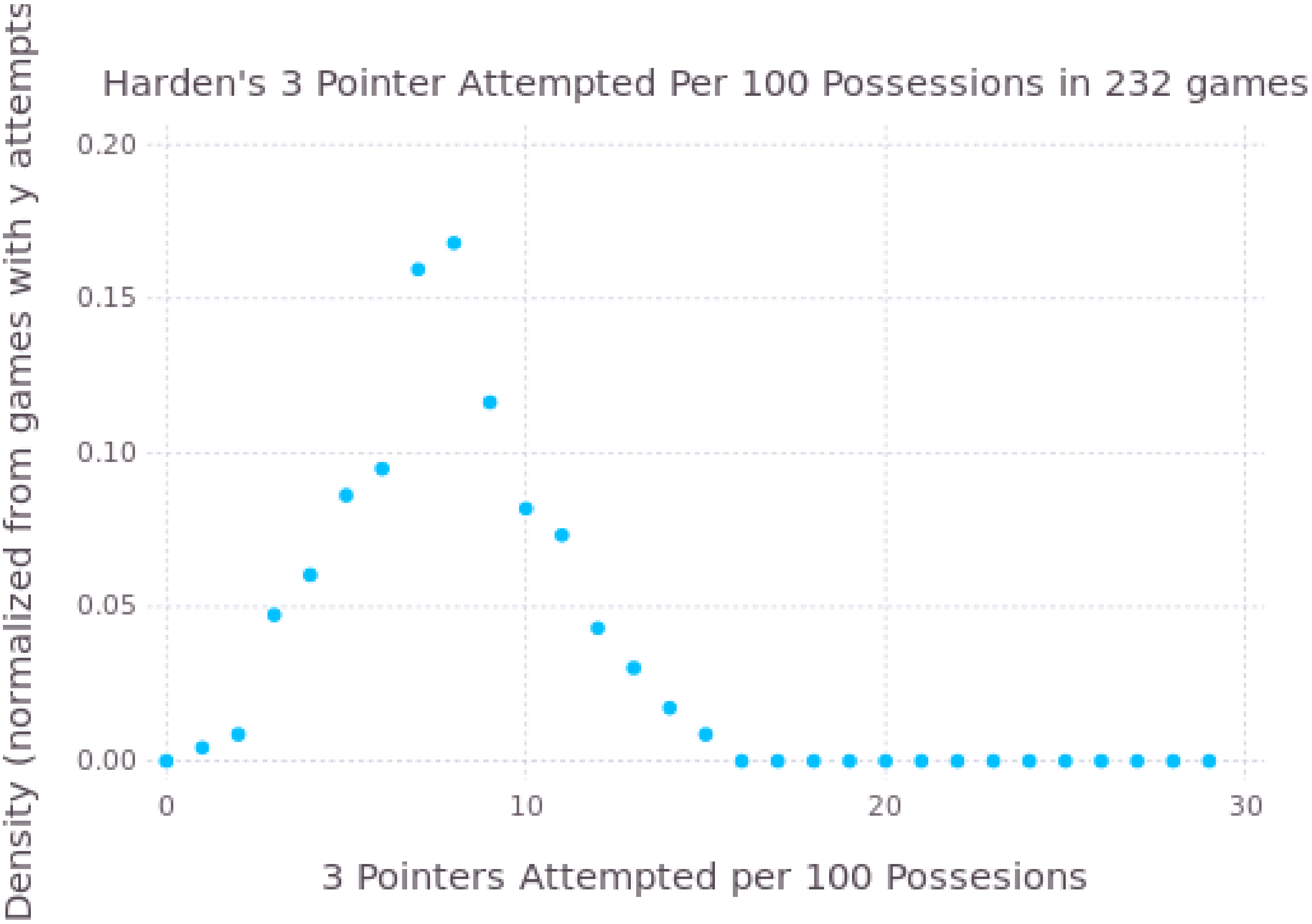
Error II

For whatever reason, there were difficulties in computing valid λ's for three-point shots attempted. Thus, there was not sufficient data to compute streakiness. Instead, I will provide the graphs of each of the scaled histograms of players' attempts. In order, the players to whom these graphs correspond to are Kobe, Curry, Korver, and Harden.


Kobe's 3 Pointer Attempted Per 100 Possessions in 177 games
Density (normalized from games with y attempts
0.20
0.15
0.10
0.05
0.00
0
10
20
30
3 Pointers Attempted per 100 Possesions


Curry's 3 Pointer Attempted Per 100 Possessions in 236 games
Density (normalized from games with y attempts
0.15
0.10
0.05
0.00
0
10
20
30
3 Pointers Attempted per 100 Possesions


Korver's 3 Pointer Attempted Per 100 Possessions in 220 games
Density (normalized from games with y attempts
0.15
0.10
0.05
0.00
0
10
20
30
3 Pointers Attempted per 100 Possesions


Harden's 3 Pointer Attempted Per 100 Possessions in 232 games
Density (normalized from games with y attempts
0.20
0.15
0.10
0.05
0.00
0
10
20
30
3 Pointers Attempted per 100 Possesions