A Study of Techniques used for *Regression to the Mean* Studies

Brian Vernarsky
(Dated: March 14, 2015)

I. INTRODUCTION

When I started to look into the traditional tools that are used to analyze basketball statistics I was surprised to see regression to the mean coming up as often as it did. My initial reaction was a visceral one: This is not what regression to the mean is supposed to be used for. It is a tool that is useful in population studies, to deal with statistics that are completely governed by chance, and not skill. Things like the average height of a randomly selected group of people in a city (the more people you measure you'll eventually get closer to the mean of the city as a whole, regardless of whether you started with a string of exceptionally tall or short people), or how well a random group of people can predict the result of fair coin flip. But it is not useful in situations where differing levels of skill, ability, genetics, or some other biasing factor, come into play. Thus, if the coin that is being flipped is not fair, and some of the people making predictions realize that and take it into account, and others don't, those people's predictions will not regress to the mean; they have a skill that the rest of the population does not.

Basketball (and all other sports) involve people whose skill levels are all well above average, but even among themselves, there is a wide variation in those skills. Assuming that a player with a hot start must cool off and will eventually start playing more like everyone else, ignores the possibility that he may be an above-average player; and vice versa for a cool start and a below-average player. Players do regress to a mean, but they each regress to their own mean, their own particular skill level over a given period of time. The trick is then to find that skill level.

In looking more closely at some specific examples, and following the derivation used for regression to the mean from *The Book* by Tom Tango, Mitchel Lichtman and Andrew Dolphin, I realized that the method used is strong, but that I would implement it differently than it appears to be used. They do appear to use the 'regression to mean' as a way of determining a player's skill level based on his current performance and the average of some population of players, generally either all players in the league or a subset of them, which is not what it generally meant by 'regression to the mean' in statistics texts. This is a good idea and I was interested in fleshing out the details. (Strong caveat here: the only methods that I know about are ones published in books or online; it is entirely possible, and most likely that people who work for NBA (or other sports leagues) teams have their own methods that are not made public. Any comments on methods are directed towards those that I have seen. *The Book* deals specifically with baseball statistics, but it is easily generalized to any sport.)

I decided to derive all of the equations myself, and come up with an implementation for how I would use it, and for what purposes. I would not refer to my method as regression to the mean, but rather making a prediction and then refining it. The basic premise is that I would like to determine a player's true skill level in a given season. Before the season starts I can make a prediction of that skill level, based off of some criteria. Once the season starts, I will get in new information, his current stats, and I can use that to make a new prediction. I now have two different predictions, based off of mutually exclusive data; both of these guesses can be utilized in making a final prediction of the player's true skill, by means of a weighted average, using the standard deviations of the two guesses as the weights. If we call the two predictions a and b, and call the standard deviations of a and b, σ_a and σ_b respectively, then the weighted average prediction, p_{pred}, would be

$$p_{pred} = \frac{\frac{1}{\sigma_a^2}a + \frac{1}{\sigma_b^2}b}{\frac{1}{\sigma_a^2} + \frac{1}{\sigma_b^2}} \ . \tag{1}$$

Getting one of the predictions is easy — the player's current stats will provide one of the guesses — the trick comes in getting the other one. The methods derived and detailed in the rest of this paper will show one method — that of defining a population and using the statistics of that group to determine a mean and standard deviation that will be used as the prediction — but any prediction could used, so long as it has a well-defined value and standard deviation.

In Section II, I will detail how to find the mean and standard deviation of a statistic in a given population, which may be different than others use. In Section III, I will discuss how to select the proper population for the purpose desired, and the effect that the choice of population can have on the results. In Section IV, I will show an example usage, trying to predict a player's shooting percentages after January 1^{st} based on his performance before January 1^{st} in that season. Finally, in Section V I will use the best method from Section IV to predict the shooting percentages for selected players in the 2015 season after January 1^{st}, 2015.

II. METHOD

If it were possible to know a player's true skill level in a particular skill, say free-throw shooting or rebounding, at any given time, then it would be possible to predict his performance over any given number of attempts to

be within a certain range. Let us call his true skill $\mathbb{P}$ and assume it remains constant for some period of time. If during that time he has a attempts to succeed (e.g. takes a shot, or is on the floor when a rebound is available), then he will be predicted to succeed on

$$\bar{s} = \mathbb{P} \times a \tag{2}$$

of those attempts. However, due to the nature of probabilities and randomness, he may not succeed exactly $\bar{s}$ times, but instead we can define a range around $\bar{s}$ of possible values for the number of successes. The probability for succeeding s times in a attempts, given a true skill of $\mathbb{P}$ is given by the binomial distribution

$$\mathcal{P}_{\mathbb{P}}(s, a) = \frac{a!}{s!(a-s)!}\mathbb{P}^s(1-\mathbb{P})^{a-s} \ . \tag{3}$$

The value of s with the highest probability is $\bar{s}$, but there is a wide range of values of s that have non-zero probability, and his performance in this particular set of a attempts could be anywhere in that range. The standard deviation,

$$\sigma_s = \sqrt{a\mathbb{P}(1-\mathbb{P})} \ , \tag{4}$$

defines how spread out the results (s, the number of successes) can be. If repeated measurements were made of exactly a attempts, with the same true skill $\mathbb{P}$, then 68% of the resulting values for s would lie within 1σ of $\bar{s}$ and 95% would lie within 2σ.

Those formulas are correct, but they presuppose knowledge of $\mathbb{P}$, which impossible to know. Instead, we must work backwards, using measured values of s successes in a attempts to infer what $\mathbb{P}$ is. Perhaps the main job that a sports statistician has is trying to figure out each player's true skill. The best guess at $\mathbb{P}$ based only on measured performance will always be s/a where all a attempts were made during a period when $\mathbb{P}$ can be reasonably assumed to be constant. This period of constancy can be taken to be a single game, or a month, or a season, or a career; it could include attempts made in practices, if accurate records were kept, or from games in any other setting; the choices of what period and which attempts should be considered depends on the context. We will call our best guess at $\mathbb{P}$ p, which will be defined by

$$p = s/a \ , \tag{5}$$

and the standard deviation of p (the range in which we should expect the true skill $\mathbb{P}$ to be) is

$$\sigma = \sqrt{\frac{p(1-p)}{a}} \ . \tag{6}$$

From the equation for the standard deviation it is clear that the greater the number of attempts, the smaller the standard deviation is, and thus the better a guess at $\mathbb{P}$ p is. Or more precisely,

$$\lim_{a\to\infty} p(a) = \mathbb{P} \ , \tag{7}$$

where p(a) simply indicates that p is a function of a.

Each of these values has an uncertainty associated with it. For p, the uncertainty is known as the standard deviation of the mean, and is defined by

$$\sigma_p = \frac{\sigma}{\sqrt{a}} = \frac{\sqrt{p(1-p)}}{a} \ ; \tag{8}$$

and for the standard deviation, the uncertainty is

$$\sigma_\sigma = \frac{1}{\sqrt{2(a-1)}} \ . \tag{9}$$

It is of course impossible for a basketball player to attempt an infinite number of shots, or be a part of an infinite number of possessions, especially all while at the same skill level. Kareem Abdul-Jabaar holds the NBA record for field goal attempts in his career at 28,307, with a shooting percentage of 0.559, which leads to a standard deviation of 0.003. That is as small as one could ever hope to get a standard deviation on a shooting percentage, but it is not even valid because his true skill most likely changed from season to season. Thus, there is a fundamental limit to how well we can determine a player's true skill, especially over shorter ranges of time than a 20-year career. So it is helpful to have other ways of guessing a player's true skill. Then, the multiple ways of predicting that true skill can be combined together to give a better prediction.

A. Predictions from a Population

One of the many ways to predict a player's true skill is to infer how well he will do based on how well players similar to him fare. The basic idea is that if you choose the population of similar players well, then a good estimate for the player's true skill will be the mean of those players' abilities, and the range in which his abilities should be found will be related to the standard deviation of the population's abilities.

We will cover how to select the right population in Section III, but for now assume that we have a proper population of players' statistics, which consists of a set of pairs of values (s_i, a_i), where the subscript i indicates that this pair is for the i^{th} player or season. These two values are all that is needed to calculate p_i and σ_i for the pair, according to Equations (5) and (6). Now we need to combine them together. Each pair will have some statistical error, σ_i, but it is important to note that each pair will also have it's own true skill, $\mathbb{P}_i$, and thus when combining them, there will be some spread in true skill. We can safely assume that the statistical errors will be distributed *normally*, and, as long as the population was chosen well, the true skill should also be distributed normally. Therefore the distribution of all of the pairs in the population should follow Gaussian statistics.

In Gaussian statistics, the probability for obtaining a value p, given a mean of $\bar{p}$ and standard deviation σ is

$$\mathcal{P}_{\bar{p},\sigma}(p) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(p-\bar{p})^2/2\sigma^2} \ . \tag{10}$$

This function is equivalent to Equation (3) for large values of a. For our purposes, that will generally be the case. So each pair has its own version of Equation (10) with its own $\bar{p}_i$ and σ_i.

When we combine all the pairs together we get a Gaussian that is appropriate for that group, with

$$\bar{p}_{pop} = \frac{\sum_{i=1}^{N} s_i}{\sum_{i=1}^{N} a_i} \ , \tag{11}$$

where N is the number of pairs in the population. The standard deviation for the population however, is not Equation (6), because that assumes that the true skill is the same for all pairs; instead the standard deviation must account for the spread in skills as well as the statistical standard deviations of each pair. We can write it as

$$\sigma_{pop} = \sqrt{\sigma^2_{skill,pop} + \sigma^2_{random,pop}} \tag{12}$$

where each of those terms will be defined below, and they may be combined this way because we assume that the two sources of error are uncorrelated.

The true skill level is what is trying to be calculated, and so the standard deviation of that skill, $\sigma_{skill,pop}$, cannot be directly determined, but it is possible to calculate it from the other two quantities, which can be measured, as

$$\sigma_{skill,pop} = \sqrt{\sigma^2_{pop} - \sigma^2_{random,pop}} \ . \tag{13}$$

B. Calculating σ_{pop}

The standard deviation of a dataset is a well-defined quantity (I will square both sides to avoid the constant square-root sign)

$$\sigma^2_{pop} = \frac{1}{N-1} \sum_{i=1}^{N} (p_i - \bar{p}_{pop})^2 \ . \tag{14}$$

However, that assumes that each term should be treated equally. Since each pair can contain a significantly different number of attempts, and the standard deviations are different as well, it makes sense to weight certain pairs more than others. Equation (14) can then be generalized to

$$\sigma^2_{pop} = \frac{1}{\sum_{i=1}^{N} w_i} \sum_{i=1}^{N} w_i (p_i - \bar{p}_{pop})^2 \ , \tag{15}$$

where w_i is the weight for the i^{th} pair. Treating this as a basic weighted average, the weights are determined to be the inverse of the square of the standard deviation of the pair, defined by (6),

$$w_i = \frac{1}{\sigma_i^2} \ . \tag{16}$$

Which leads to a final formula for calculating the standard deviation of the dataset of

$$\sigma^2_{pop} = \frac{1}{\sum_{i=1}^{N} \frac{1}{\sigma_i^2}} \sum_{i=1}^{N} \frac{1}{\sigma_i^2} (p_i - \bar{p}_{pop})^2 \ , \tag{17}$$

C. Calculating $\sigma_{random,pop}$

The standard deviation due to statistical fluctuations about the mean of the dataset cannot be determined from binomial statistics as in Equation (6) because that equation assumes that the skill level is constant across all measurements, which is not the case here. It is determined by a weighted average of the standard deviations for each pair, using the uncertainty of the standard deviation (see Equation (9)) as the weights. Thus,

$$\sigma^2_{random,pop} = \frac{1}{\sum_{i=1}^{N} \frac{1}{\sigma^2_{\sigma,i}}} \sum_{i=1}^{N} \frac{1}{\sigma^2_{\sigma,i}} \sigma_i^2 \ . \tag{18}$$

D. Summary

Using Equations (17) and (18), it is now possible to determine the standard deviation in the skill level of the dataset from Equation (13). Now we can make a prediction of a range for the player's skill, which is

$$p_{prediction} = \bar{p}_{pop} \pm \sigma_{skill,pop} \ . \tag{19}$$

Thus we can see that the prediction is going to be very strongly influenced by the population that is chosen. How well the prediction will do will be based on $\bar{p}_{pop}$, which requires choosing players that are in fact similar to the player in question; how tightly that value can be constrained depends on how similarly those players perform. Choosing the right population is very important and requires careful forethought.

III. SELECTING THE POPULATION

For the purposes of this paper, we are considering how to predict an NBA player's true skill level in any of a number of different statistical categories: field-goal shooting percentage; free-throw shooting percentage; rebounding percentage; etc. But before we select a population to use, we must first consider the time-frame for

which the prediction needs to be valid. Is it a prediction for the next game? The next month? The playoffs? The next season? Or his entire career? Each of these possibilities may require a different population to make a reasonable prediction. Recall that skill levels for a given player will change over time, so that often it can only be assumed that he has a constant skill level over the course of a single season. Even that may be a stretch, but, barring an injury that forces him to miss significant time in the middle of a season, it is a reasonable assumption. It may be useful to use longer stretches than a single season, say 150 games, or two seasons, so in general we can call these chunks of games *gamesets*.

The gamesets can start and end at any point in time, but it is generally preferrable for the games to all be contiguous (i.e we consider all games between two dates without excluding any), but there are sometimes reasons for excluding a game (e.g. the player didn't have any attempts for the stat in question that game, or he was injured at the opening tip-off). All games in a single season would be one example of a gameset; sets of 50 games would be another. If a player played 250 games over 5 seasons, he would have 5 gamesets of seasons, and 5 gamesets of 50 games, but they may include different games. The size of the gameset should be proportional to the time-frame the prediction needs to be valid. The prediction for a season come from a population made up of gamesets of seasons, or chunks of 50-100 games, but shouldn't include datasets of 500 games or entire careers.

The next choice for the population is what players should be included. Since we are only considering NBA players, a good first choice at a population would be all NBA players. It is clear that by lumping all NBA players together the standard deviation will be quite large, because we are making no attempt finding similar players, other than having played in the NBA. Another choice would be all players of the same position (point guard, shooting guard, small forward, power forward, or center). This will decrease the standard deviations slightly; this attempts to separate players by skill as in general guards are more similar to other guards than they are to centers. However, it may create complications for players who play multiple positions, and players who don't play like other players at their position (e.g. a center who regularly shoots three-pointers). Another option is to use some metric to determine which players are most similar to the player in question; this will be discussed further below.

When making a prediction we can only include games that have come before the prediction (this is obvious when making predictions about the present, but when make predictions for the past to see if the method works, it is a necessary requirement). But should we include all games by all players in all seasons prior to the prediction? Given the changes in rules (shot-clock, three-point line, restrictions on styles of defense, etc), style of play, and overall talent that happens over time, it seems wise to restrict the pool of games to choose from to some number of recent seasons. Using only the previous season as the pool to select from would be reasonable, but including more seasons increases the statistics and leads to better results; however, including too many seasons may change the skill level too much, or ignore newer rule changes. Five seasons seems like a reasonable number of previous seasons to include when considering all NBA players, or all NBA players of a certain position. If using only select gamesets from select players (i.e. when using only a small number of similar player as determined from some metric), then it might be useful to consider further back than just five years if not enough similar players are found. If considering a stat like free-throw shooting percentage, then perhaps a wider time-frame could be used, since there have been no significant rule changes that would affect that percentage. When considering a player's own past performance, then all games in his career should be allowed to be used, and how many are actually used will depend on the context.

Thus, when selecting the population to use to make the predictions from Section II, it is important to consider the time-frame that you are interested in predicting, and therefore the size of the gamesets; the players which should be included, whether all players, a subset of players, or the previous performance of the player in question himeself; and finally how far back in time you're interested in going. The answers to each of these questions depend on the context of the prediction and how precise you want to be.

A. Distance Metric

The idea of using a metric to determine which players should be in the population was mentioned in the last section. I won't go into exhaustive detail here, because it was already discussed in my previous work, but I will repeat the formula here, as I have modified it for the current purpose. The basic thought is to define a *distance* between two different data points, in this case seasons. We are interested in finding seasons that are *close* to our season in question. Thus, we define the distance metric as

$$d_{ij}^2 = \sum_{k=1}^{4} w_k [\frac{s_k^i - \bar{s}_k^{i,pop}}{\sigma_k^{i,pop}} - \frac{s_k^j - \bar{s}_k^{j,pop}}{\sigma_k^{j,pop}}]^2 , \qquad (20)$$

where d_{ij} is the distance between seasons i and j, with season i being the season in question; the s_k are the statistics we will be using; $\bar{s}_k^{i,pop}$ is the mean of s_k for the population from which season i was drawn (i.e. if season i is from the 2014 season, then $\bar{s}_k^{i,pop}$ is the mean of s_k for all players in the 2014 season); $\sigma_k^{i,pop}$ is likewise the standard deviation for the population from which season i was drawn; and the w_k are the weights assigned to those statistics.

Equation (20) can be used with any number of different statistics, so long as they have well-defined means and

standard deviations. Given a season that you would like to find the N most similar seasons, it is only necessary to search through all the possible player-seasons available (this may be limited depending on the objective) and calculate d_{ij} for each pair, and then select the N closest.

IV. EXAMPLE: SHOOTING PERCENTAGES

Let us take as an example predicting three different shooting percentages for a player from January 1^{st} until the end of the season, given his shooting percentages before January 1^{st} of that season. The three different shooting percentages are two-point shooting, three-point shooting, and free-throws. I am using three different shooting percentages to show how different stats need to be treated different. Since we are making our predictions on January 1^{st}, we already have a predictions of the true skill levels for this player in this season, which are just his current shooting percentages, with their standard deviations. But we can come up with a better guess for his true skill level by defining a population, obtaining a mean and standard deviation, and then applying Equation (1). The question is: which population should be used to generate the mean?

A. Populations Used

We will try out several different populations and see how well they each perform. The populations are listed below. In all populations, only players from seasons after 1985 are considered, and the player had to play at least 100 minutes (just over two full games worth) in the season.

- The full-season results for all players in the NBA over the previous five NBA seasons. Each year from 1990-2015 will have its own mean and standard deviation that is derived from the five seasons preceeding it. The means and standard deviations are derived using Equations (11) and (13).

- The full-season results for all players in the NBA over the previous five NBA seasons, broken down by the five different positions. Each year and each position from 1990-2015 will have its own mean and standard deviation derived from Equations (11) and (13).

- The last 50, 60, 70, ..., 150 games, as well as career totals, for the player in question. Each gameset will be treated separately, and includes the last X number of games the player played before the season in question started. The mean and standard deviations are derived using the binomial formulas in Equations (5) and (6). In order for this prediction to be made, the player must have played X number of games so far in his career before the start of the season. If he has played less than that many games, no prediction will be made using this method.

- The next season results of the 100 *most similar* players to the player in question, as defined in Section III A, from the five seasons prior to his last season. Thus, if predicting for the 2015 season, his 2014 season results would be used to find similar players in the 2009-2013 NBA seasons. The statistics used in Equation (20) are the percent of possible minutes played in the season ($\frac{minutes\ played}{48*games\ played}$), the number of shots of the type under study (two, three, or free throw) attempted per minute played ($\frac{shots\ attempted}{minutes\ played}$), and the shooting percentage for the type of shot under study; the weights are 0.25, 0.25, and 0.5, respectively. The means and standard deviations of these next season results are derived using Equations (11) and (13). In order for this prediction to be made, the player must have played at least 100 minutes in the season immediately preceeding the one in question, and attempted a shot of the type under study at least once every 100 minutes of game time. For seasons to be considered *close* to the season in question, the player must have played at least 100 minutes in the season in question and also the season immediately following it. Note that this method cannot be used for rookie seasons.

Each of these methods will be applied to every season of every player who plays at least 20 games before January 1^{st} of that season with at least one shot attempt of the type under study, and at least 20 games after January 1^{st} of that season with at least one shot attempt of the type under study. These restrictions are set to ensure that there are enough attempts both before and after January 1^{st} to have a reasonable shooting percentage to make predictions and comparisons. By default, the strike-shortened 1999 and 2011 seasons are excluded as no players had more than 20 games before January 1^{st} in those years.

B. Assessing the Quality of the Predictions: χ^2

To determine how well the predictions do, the predicted shooting percentages will be compared to the actual shooting percentages after January 1^{st}, taking into account the standard deviations of each. This is done using the χ^2 method. χ^2 is computed as

$$\chi^2 = \sum_{players} \frac{(p_{pred}^{player} - p_{meas}^{player})^2}{\sigma_{p_{pred}}^2 + \sigma_{p_{meas}}^2} \ , \tag{21}$$

where the sum is over all player-seasons that a prediction was made for, p_{pred} is the predicted shooting percentage, $\sigma_{p_{pred}}$ is the standard deviation of that percentage, p_{meas} is the measured shooting percentage, and $\sigma_{p_{meas}}$ is the

standard deviation of that percentage calculated using Equation (6). The reduced χ^2 is just χ^2/NDF, where NDF is the number of degrees of freedom, which in this case is just the number of seasons analyzed minus 1.

A reduced χ^2 of near 1 is considered optimal. This essentially means that the standard deviations associated with the predicted means are commensurate with the actual differences seen in the data. If the reduced χ^2 is near 1 for several of the predictions, then the result with the smaller standard deviation is considered better, since it gives a more precise prediction; although it is also important to note the number of games for which a prediction can be made using that method. For instance, using the previous 1,000 games for a player to make a prediction may give an excellent prediction, but it would only be possible for a handful of players, and so it is not as useful as one that can be used for a greater number of players.

C. Data

The data used in this analysis were obtained from the Basketball Reference website, which has the statistics for every professional basketball player in NBA and ABA history. The game-by-game stats were obtained for every game played by all players whose careers ended after 1985. Thus, for a player whose 10-year career ended in 1982, all of his games were obtained, even those from before 1985. This was done to ensure career statistics were able to be easily generated. For some seasons before 1985, certain statistics are not recorded, for instance the number of field goals attempted for each game, only the number of made baskets was available, however the full-season stats included the total number of attempts. This is why 1985 was used as a cut-off.

The cuts placed on the data were mentioned when describing each of the populations used. The position played by a player is not recorded on a game-by-game basis in the statistics obtained, but only in the season summary. Players who are listed as playing multiple positions during a season have their full stats added to both positions when separating players by position. This is done because it is unknown how many games were played at each position, and it is possible that during certain games he may even play multiple positions, so it would be very difficult to ascertain how to properly split up the statistics. In making predictions for his skill, the position listed first was used, as it is considered his primary position. Strike-shortened seasons (1999, 2011) are not included in the study, but are included in career statistics and groups of the last X (50, 60, ..., 150) games.

D. Results

The results of the different test populations can be seen in Table I for two-point shooting percentages, in Table II for three-point shooting percentages, and in Table III for free-throw shooting percentages. The tables have 17 different populations, the ones I've mentioned so far, as well as well as a second entry for League and Position entries, indicated with a (d). These two extra entries are the values for the League and Position predictions only in games in which a prediction was made using the distance method; this just allows us to compare apples to apples.

It should be noted here that the standard deviations for the League and Position populations will by definition be larger than any other subset of players. Essentially what that prediction does is to predict the exact same shooting percentage and standard deviation for every player in the league, or every player in the league at a given position, in a given season. It has no *a priori* bias for any player. It is a fairly simple prediction, just assume everyone will shoot at the league average for the last five years, and the spread should be roughly the same as it was over the last five years. It is almost guaranteed to be right, it's just not terribly useful, because it treats all players the same and thus has a large spread. It's main usefulness is that it can be applied to all players and seasons.

The standard deviations for the Last X number of games and career populations are naturally much smaller, and get smaller as the number of games increases. This is because I am using only the binomial standard deviation, which gets smaller as more attempts are added. It does not take into account skill levels varying with time. It would be necessary to come up with a way to account for this before putting these to more use. It's main benefit is that it uses the player's own statistics to generate a prediction. A drawback though is that it can only be used after a player has played a certain number of games in his career, which means it cannot make predictions for all players.

The distance method combines the benefits of the other two general methods: it has well-defined standard deviations, that are appropriate for the average differences seen, and it is personalized to each player. It still cannot make predictions for all possible seasons, as it requires that the player have played in the previous season, but it still is able to make predictions for nearly 90% of the possible seasons. The fact that such a simple distance metric, including only minutes played per game, shots attempted per minute played, and shooting percentage, performs so well is encouraging. In the Conclusion I will discuss ways to make it better.

1. *Two-Point Shooting Percentages*

Looking at the two-point shooting percentage results in Table I, we can see that the League, Position, and Distance populations have very good values of χ^2/NDF, very close to 1. This indicates that the analysis done in this paper has been done correctly. The method is projecting reasonable means, and the standard deviations do a good job of accounting for the spread seen in the data. The Last X populations have slightly higher

Population	χ^2/NDF	# Seasons	avg diff	avg $\sigma_{pop,pred}$
Last 50	1.202	4377	0.032	0.022
Last 60	1.222	4292	0.031	0.020
Last 70	1.244	4171	0.031	0.019
Last 80	1.267	4003	0.031	0.018
Last 90	1.284	3886	0.031	0.017
Last 100	1.307	3829	0.031	0.016
Last 110	1.324	3779	0.030	0.016
Last 120	1.338	3707	0.030	0.015
Last 130	1.354	3627	0.030	0.015
Last 140	1.365	3546	0.030	0.014
Last 150	1.387	3439	0.030	0.014
Career	1.618	5039	0.033	0.013
League	1.084	5039	0.033	0.026
Position	1.069	5039	0.032	0.025
Distance	1.102	4461	0.030	0.021
League(d)	1.090	4461	0.033	0.026
Position(d)	1.076	4461	0.032	0.025

TABLE I. Results for the two-point shooting percentage tests. Note that the number of games is different for most of the populations. The top part of the table uses the player's own games as the population, while the bottom part of the table uses some subset of the league in the last five years as the population. The League(d) and Position(d) entries are there to illustrate how well the League and Position population predictions do only in seasons for which a prediction can be made using the distance method. The avg diff column is the average absolute value of the difference between the predicted result and the measured result. The avg $\sigma_{pop,pred}$ column is the average standard deviation for the prediction coming from the population.

Population	χ^2/NDF	# Seasons	avg diff	avg $\sigma_{pop,pred}$
Last 50	1.129	1894	0.043	0.030
Last 60	1.141	1840	0.042	0.029
Last 70	1.138	1766	0.041	0.027
Last 80	1.149	1703	0.041	0.026
Last 90	1.156	1665	0.040	0.024
Last 100	1.177	1625	0.040	0.023
Last 110	1.186	1588	0.039	0.023
Last 120	1.195	1547	0.039	0.022
Last 130	1.192	1508	0.039	0.021
Last 140	1.210	1457	0.039	0.020
Last 150	1.210	1409	0.039	0.020
Career	1.358	2164	0.042	0.019
League	1.040	2164	0.044	0.035
Position	1.042	2164	0.043	0.034
Distance	1.038	1977	0.041	0.029
League(d)	1.022	1977	0.043	0.034
Position(d)	1.023	1977	0.043	0.034

TABLE II. Results for the three-point shooting percentage tests. See Table I for a description of the fields.

χ^2/NDF values, indicating that the standard deviations have been underestimated. This is somewhat expected, as I used the binomial standard deviations and made no attempt to account for varying skill levels. Within 50 games this shouldn't cause such an issue, but over 150 games it certainly might.

All of the average differences between the predicted and measured values fall in a range from 0.030-0.033, while the standard deviations show a bigger range. The distance calculation seems to be the best, with a reasonable χ^2/NDF value, the smallest average difference among the group and standard deviations that are well below those of those of the other league-wide derived averages.

2. Three-point Shooting Percentage

Again, the League, Position, and Distance metrics have significantly better χ^2/NDF values than the Last X games populations, however those are improved here as well, perhaps indicating that three-point shooting skill changes less over time than two-point shooting skill.

Here, it is clear that the distance method gives the superior results. The average difference is a little bit less, and the average standard deviation is a full percentage point lower than the other two methods. Note that there are significantly fewer seasons included because fewer players shoot three-point shots with regularity.

3. Free-Throw Shooting Percentages

Interestingly, the free-throw shooting percentage proves to be the most difficult to predict. The χ^2/NDF values for all populations are worse here than for the other shooting percentages. I think the reason this happens is that the distribution for free-throw shooting percentages are not distributed *normally*. Figure 1 shows the distribution of the shooting percentages (weighted by their standard deviations) for each of the three different kinds of shot for the years 2009-2014. The line on each figure is a Gaussian function with the mean and standard deviation as calculated from Equations (11) and (13). The two- and three-point shooting percentage distributions are very well-matched by the Gaussian approximation, but the free-throw shooting percentage is not, largely because the mean is so close to 1. Since it is not possible to record a percentage over 1, the Gaussian is therefore asymmetrical, and the mean is lower than it should be. In order to improve these predictions it will be necessary to find a different way to define a mean and standard deviation for these points. This can be accom-

Population	χ^2/NDF	# Seasons	avg diff	avg $\sigma_{pop,pred}$
Last 50	1.537	2171	0.038	0.022
Last 60	1.607	2116	0.038	0.021
Last 70	1.644	2045	0.038	0.020
Last 80	1.684	1973	0.037	0.019
Last 90	1.714	1930	0.037	0.018
Last 100	1.763	1887	0.037	0.017
Last 110	1.793	1834	0.037	0.016
Last 120	1.829	1794	0.037	0.016
Last 130	1.857	1746	0.036	0.015
Last 140	1.894	1702	0.036	0.015
Last 150	1.938	1649	0.036	0.014
Career	2.450	2468	0.041	0.014
League	1.234	2468	0.043	0.036
Position	1.240	2468	0.042	0.035
Distance	1.183	2294	0.038	0.030
League(d)	1.216	2294	0.042	0.036
Position(d)	1.222	2294	0.041	0.034

TABLE III. Results for the free-throw shooting percentage tests. See Table I for a description of the fields.

plished with the use of fitting functions to determine the mean and employing asymmetrical standard deviations, but that's another rabbit hole entirely, and the subject of another paper. That being said, the distance method again provides the best average differential and average standard deviation, however it is clear that the current method is not all that accurate.

E. Discussion

From the predictions that have been presented in this section, it is clear that the method we have outlined is a robust method, producing reliable estimates for a player's mean shooting ability, while also providing a good range around that mean in which we should reliably be able to expect to find his actual results. The only real caveat is what we saw from the free-throw shooting predictions: we require that the shooting abilities be distributed *normally*, which may not always be the case. In those cases, another method of defining a mean and standard deviation must be developed and utilized. As I mentioned, using asymmetric standard deviations results in much better agreement.

It is also apparent that the distance metric does a fairly good job at picking *similar* players. Since the same basic method is used to determine the mean and standard deviation for the prediction here as with the League and Position methods, albeit with significantly fewer entries, it stands to reason that those values are being computed correctly. Thus, we should expect a χ^2/NDF of near 1, so long as we have chosen our population well. It is apparent from the χ^2/NDF values for the two- and three-point shooting percentages that populations have been chosen well. The advantage here is that the average standard deviation has gone down by almost a full percent in both instances, which is a huge improvement in precision. I again point out that this is a very simple method at the moment, with only three parameters. Introducing more parameters should lead to better results, although there is a limit to how well anything can be predicted.

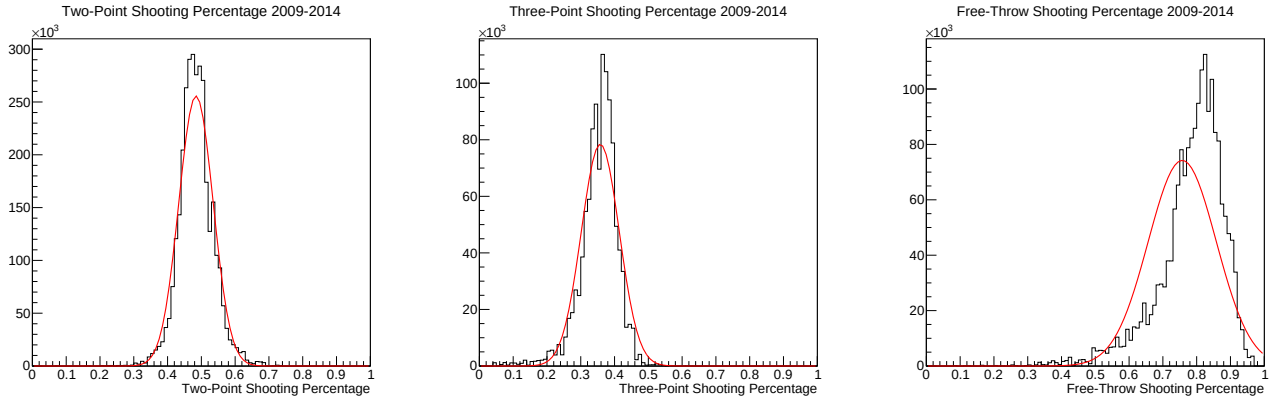
Going forward, it seems a reasonable plan to use the distance method to make any predictions where there exists a prior season to use to find similar players; and where that is not available, it is acceptable to use the broader League-wide averages.

V. 2015 PREDICTIONS FOR 76ERS

A useful prediction at the moment is: how well will the some current NBA players do in the 2015 NBA season. I have thus chosen to look at the current 76ers roster, and make predictions on their shooting abilities for the rest of the 2015 season. Predictions are only made for players who attempted at least 1 shot of each kind in 20 games before January 1^{st}, 2015. The results can be seen in Table IV. Note that Robert Covington, K.J. McDaniels, and Nerlens Noel are in their first qualifying years (over 100 minutes played), and so don't have distance method predictions (Also Note: It was very interesting to me to see that Robert Covington has a prediction for three-point shooting, but not two-point shooting. It seems he only attempted a two-point shot in 19 games, while he attempted a three-point shot in 20 games. The game he shot only three-pointers? December 12, 2014 against Brooklyn, when he made 6/10 threes. I guess it makes sense why he only shot threes that game!)

VI. CONCLUSION

The results of this experiment have been very positive. The equations derived for the predictions are different than those used elsewhere, but stand up to the reduced χ^2 test. I used a different set of data than others might use in making my predictions, and different methods, and still came up with very reasonable results. I believe that my method is stronger than a method that relies on how other players are doing in the current season to make a prediction. Since the league does not change significantly over a 5-year period, it is justifiable to use that data to make predictions for the year immediately following that period. Using five seasons worth of full-year data makes it much easier and more reliable to determine the league average and the standard deviation in skill. Plus, there is no danger of biasing the data with the player you are trying to study's own stats. Thus, I am very comfortable with these predictions, but wouldn't refer to my method



(a) (b) (c)

FIG. 1. The shooting distributions from 2009-2014 for (a) two-point, (b) three-point, and (c) free-throw percentages, weighted by the standard deviations of the entries. The line on each figure is a Gaussian with the measured mean and standard deviation. Note that (a) and (b) are well-defined by the Gaussian, while (c) is asymmetric, and the agreement is much worse.

Name	p_{league}	σ_{league}	$p_{distance}$	$\sigma_{distance}$
Two-Point Shooting Percentage				
Michael Carter-Williams	0.436	0.026	0.445	0.020
Luc Mbah a Moute	0.475	0.031	0.469	0.027
K.J. McDaniels	0.483	0.033	N/A	N/A
Nerlens Noel	0.455	0.029	N/A	N/A
Henry Sims	0.506	0.029	0.496	0.020
Hollis Thompson	0.476	0.038	0.485	0.028
Tony Wroten	0.480	0.029	0.481	0.021
Three-Point Shooting Percentage				
Michael Carter-Williams	0.298	0.038	0.314	0.026
Robert Covington	0.399	0.037	N/A	N/A
Luc Mbah a Moute	0.307	0.038	0.297	0.038
K.J. McDaniels	0.325	0.037	N/A	N/A
Hollis Thompson	0.352	0.037	0.361	0.025
Tony Wroten	0.314	0.036	0.316	0.029
Free Throw Shooting Percentage				
Michael Carter-Williams	0.651	0.043	0.665	0.038
K.J. McDaniels	0.763	0.046	N/A	N/A
Henry Sims	0.770	0.047	0.773	0.038
Tony Wroten	0.673	0.038	0.662	0.036

TABLE IV. Predictions for the shooting percentages after January 1^{st}, 2015 for 76ers players who have attempted a shot of each type in at least 20 games before January 1^{st}, 2015. Two predictions are given, one using the league averages and mean, and the other using the distance method. Players with an N/A in the distance columns do not have a distance prediction because they did not play enough in the 2014 season.

of predicting a player's future performance as *regressing to the mean*, rather I would call it *refining a prediction*.

The distance metric used in this paper is infinitely extendable. Any combination of statistics can be used to determine *similar* players. The completely bare-bones set used in this paper produced excellent results, but that can be improved with other statistics. For instance: only single seasons were used here, but two or three seasons could be used, or career data; a cut could be placed on age, such that only players within 5 years of the player's age would be considered; offensive rebounding could be included for two-point shooting percentages, as put-backs are a common way of getting more two-point shots; the ratio of two- and three-point shots to the total number of shots could be used; etc. Also, while it was only used for shooting percentages here, it can easily be applied to any kind of percentage, rebounding, assists/possession, etc, so long as they are distributed normally.

While what was done in this paper was done uniformly for all players in the league meeting a certain criteria, the approach can and should be tailor-made for each player individually, and for the purpose for which you want to use the prediction. Once the basic formulas are set up, they can serve any number of purposes.