Finding an Ultimate Limit for an NBA Player's Shooting Percentage

Tom Kennon[*]

27 April 2018

Abstract

Extreme value theory is used to estimate the ultimate upper or lower limit for an NBA season's league leading player's shooting percentage (free throw, 2 point field goal, and 3 point field goal). The limits are found using the generalized extreme value distribution with parameters optimized using the Nelder-Mead method maximizing the loglikelihood. Two different techniques are applied in this project to finding an optimal generalized extreme value distribution location parameter μ including a constant and a Gompertz curve. A 95% bootstrap confidence interval is calculated for these limits. Kolmogorov-Smirnov tests and a Score test are run as well through the bootstrapped datasets to evaluate goodness of fits.

Keywords: NBA, Shooting Percentage, Extreme Value Theory, Endpoint Estimation

Contents

1 Introduction **2**

2 Data **3**

3 Methodology **4**

3.1 Distribution 4
3.2 Fitting GEV/GEVr Location Parameter 5
3.3 Optimizing GEVr Parameters 6
3.4 95% Parametric Bootstrapped Confidence Interval 7
3.5 Goodness of Fit Testing 7

4 Results **8**

4.1 Free Throw 8
4.2 3 Point Field Goal Percentage 12
4.3 2 Point Field Goal Percentage 13

5 Discussion **14**

[*] thomas.kennon@uconn.edu; Undergraduate Statistics student at University of Connecticut.

1 Introduction

National Basketball Association (NBA) teams in recent years have seen an influx of usage of advanced statistical metrics to track and evaluate both players and strategy. Increasingly, NBA teams have been vehemently searching for efficiency to perform the best. NBA teams are now constantly on the look out for the most efficient shooters in the league, and thus the question of "what is the best or worst conceivable shooting percentage to lead the league?" is imperative. The three shooting percentages I will be investigating are free throws, two point field goals, and three point field goals. Free throw percentage will be the main statistic where I have outlined my methodology and I extend this methodology to two point field goal percentage and three point percentage as well.

Typically in the NBA, the league leader in free throw percentage for a season is approximately 91-93%. The highest mark to lead the league was Jose Calderon's 98.1% in 2008-2009. Calderon actually had a remarkable streak of 87 consecutive free throws during this record-breaking season. The lowest mark to lead the league was M. Zaslofsky 's 84.3% in 1949-1950. The league leader in 2pt field goal percentage for a season is typically approximately 65-70%. The highest mark to lead the league was Wilt Chamberlain's 72.7% in 1972-1973. The lowest mark to lead the league was N. Johnston's 44.7% in 1956-1957. Typically, the league leader in three point field goal percentage for a season is approximately 45-50%. The highest mark to lead the league was Kyle Korver's 53.6% in 2009-2010. The lowest mark to lead the league was Bruce Bowen's 44.1% in 2002-2003.

Free throws are one of the most unique opportunities in sports where a player is given a "free" uncontested attempt to score. Normally in a basketball game when players shoot there are defenders trying to block their shot attempts. If a defender illegally makes contact with the shooter, then the player is sent to the free throw line to attempt free throws. The exact number of free throws allotted depends on the game situation but fluctuates among 1,2, and 3. The shooter is allowed to take their time during their free throw attempts and no opposing player is allowed to defend these attempts. Ideally, a player should make every free chance they get by shooting 100% of their free throws, yet no NBA player has ever completed this feat for a whole season in the 70+ year history of the league. The NBA requires that for a player to qualify as a league leader they must meet a minimum of 125 made free throws for the season (approx. 1.5 per game). This ensures that a player cannot lead the league in free throw percentage by only taking a few shots. With NBA players shooting better and better each season as of recently, a burning question for many NBA fans is "is it possible to achieve a perfect free throw season making every shot one takes?"

Free throw shooting is important, but it is only a small part of the game of basketball. Scoring in regulation includes two point and three point shots as well. Any shot within the three point arc is considered a two point field goal attempt, and conversely any shot outside the three point arc is considered a three point field goal attempt. These shots are usually heavily contested however and are thus made successfully at lower percentages than free throws. The league leading three point percentages are generally lower than the league leading two point percentages as one would expect because three point shot attempts are taken a further distance away from the basket (behind the three point line). To qualify as a league leader in the NBA for two point field goal percentage, a player must successfully make a minimum of 300 made two point field goals for the season (approx. 3.5 per game). Similarly, to qualify as a league leader in the NBA for three point field goal percentage, a player must meet a minimum of 82 made three point field goals for the season (approx. 1 per game).

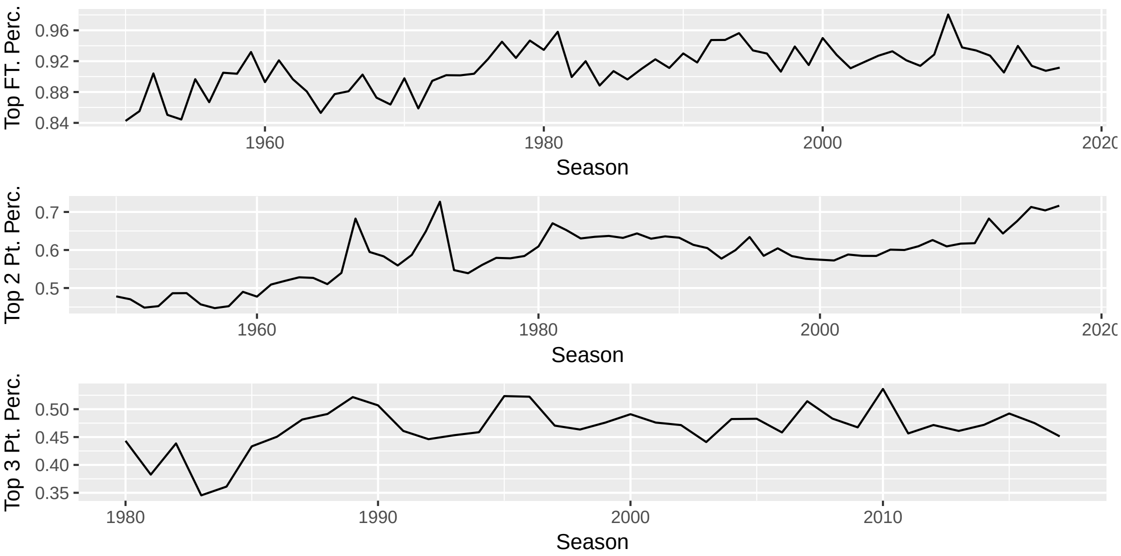

Figure 1: NBA Top Free Throw, 2 Point Field Goal, and 3 Point Field Goal Percentage Time Series Plots.

The generalized extreme value (GEV) distribution is used to estimate the upper limit because it ensures that our limit is not dependent on time as would be prevalent in other time series or extrapolation methods. This distribution requires continuous data. Much existing literature has used this approach to various sports. Einmahl and Magnus (2008) predicts the ultimate world-record limits for various track events for both the men's and women's record. For the running events (eg. 100m, 200m, etc.) a lower limit for the times is reported, and for the throwing and jumping events an upper limit for the distances is reported. Einmahl and Smeets (2011) refines this method applied to an updated dataset for the 100m dash world record. Chiou et al. (2015) applies end point estimation with the GEV distribution for baseball analysis to estimate the top baseball batting average for an MLB season. There is not much existing literature covering extreme value analysis with basketball statistics and more specifically, shooting percentages. Bader et al. (2017) extends the GEV distribution to the GEVr distribution which includes an r value that is the number of top performers included (i.e. using top 1, top2, ... or top 10 to estimate the ultimate limit. Bader et al. (2017) also outlined a procedure for selecting this optimal r to maximize accuracy of the estimate, while also maintaining a smaller interval for the predicted ultimate limit. The R package *eva* (Bader, 2016) integrates some useful functions for finding this optimal r. This package also contains functions to estimate the optimal parameters for a GEVr model based off of data if the location parameter is a constant. Marcos and Woodworth (2017) uses this package to analyze extreme sea levels.

2 Data

Figure 1 contains time series plots generated by the R package *ggplot2* (Wickham and Chang, 2016) and data wrangling performed using the R package *dplyr* (Wickham et al., 2017). The dataset is from Basketball Reference (LLC, 2018). This website is a comprehensive source for various statistics in basketball's history, most specifically the NBA. The datasets consist of the top 10 NBA individual players' free throw percentages, top 10 NBA individual players' 2 point field goal percentage, and top 10 NBA individual players' 3 point percentages respectively, for each season recorded since the inception of the league 1949-2017 (68 seasons total).

As observed by the first plot in Figure 1, the league leading free throw percentage appears to increase over time, but not in a classical linear pattern. There is an "S" curve shape to this data that appears to have two relatively stationary tails. I have taken two approaches to model this data. I fitted a Gompertz curve for the location of the curve for the entire dataset due to its similar "S" curvature and asymptotic behavior. I then also separately fitted a constant for the location of the curve for the upper stationary tail beginning after 1976.

The second plot in Figure 1 displays the league leaders in 2pt field goal percentages. There is a clear increase in the percentage over time. This data similarly has a Gompertz curve fitted to it due to the increase over time and tails behavior.

The three point data shown in the third plot in Figure 1 begins at 1980 due to the fact that this is when the three point shot was first introduced into the NBA. The NBA players had to adjust to this new opportunity and learn how to best use it to their advantage which can explain the large variability and lower average in the first 6 years of the dataset. Because the rest of this dataset is relatively stationary, using only the 7th year and beyond, a constant was fitted for the location of the curve.

3 Methodology

3.1 Distribution

3.1.1 Generalized Extreme Value Distribution

The generalized extreme value distribution is a popular distribution used in many extreme value analysis literature to predict ultimate limits as discussed in Section 1. Let x_i denote for i's season, the 1st best shooting percentage for $i = 1, 2, 3, ...68$ (68 total seasons). The GEV distribution only considers the league leading percentage (top 1) for each season. The GEV distribution takes three parameters: location $\mu \in \mathbb{R}$, scale $\sigma > 0$, and shape $\xi \in \mathbb{R}$. The GEV distribution's probability density function is defined as:

$$f(x|\mu,\sigma,\xi) = \frac{1}{\sigma}\left(1+\xi\frac{x-\mu}{\sigma}\right)^{-\frac{1}{\xi}-1} e^{-\left(1+\xi\frac{x-\mu}{\sigma}\right)^{-\frac{1}{\xi}}}, \tag{1}$$

with support:

$$x \in \left\{ \begin{array}{ll} [\mu - \frac{\sigma}{|\xi|}, \infty), & \text{for } \xi > 0 \\ (-\infty, \infty), & \text{for } \xi = 0 \\ (-\infty, \mu + \frac{\sigma}{|\xi|}], & \text{for } \xi < 0 \end{array} \right\}.$$

The support for this distribution is key in estimating an upper or lower endpoint limit. If the shape parameter is positive ($\xi > 0$), then an ultimate lower limit $x^* = \mu - \sigma/|\xi|$ can be calculated. If the shape parameter equals 0 ($\xi = 0$) then no limit can be can be calculated with this distribution. And lastly, if the shape parameter is negative ($\xi < 0$), then an ultimate upper limit $x^* = \mu + \sigma/|\xi|$ can be calculated. The location parameter μ does not have to be restricted to only one constant value. This allows for flexibility in it's prediction power. I implemented and evaluated the fit of two different location parameters including an evaluated Gompertz curve and a constant value. An extension of this generalized extreme value distribution is the generalized extreme value r distribution. The GEV distribution is a special case of the GEVr distribution where $r = 1$.

3.1.2 Generalized Extreme Value r Distribution (GEVr)

Let $x_{r,i}$ denote for i's season, the rth best free throw percentage. Let $x_{1,i} \geq x_{2,i} \geq x_{3,i} ... \geq x_{10,i}$ for $i = 1, 2, 3, ...68$ (68 total seasons). The GEVr distribution's probability density function (Bader et al., 2017) is defined as:

$$f_r(x_1, x_2, ..., x_r|\mu, \sigma, \xi) = \sigma^{-r} \exp\left\{-(1+\xi z_r)^{-\frac{1}{\xi}} - \left(\frac{1}{\xi}+1\right)\sum_{j=1}^{r} \log(1+\xi z_j)\right\}, \tag{2}$$

where $x_1 > \cdots > x_r$, $z_j = (x_j - \mu)/\sigma$, and $1 + \xi z_j > 0$ for $j = 1, \ldots, r$.
This has the same support as GEV:

$$x \in \left\{ \begin{array}{ll} [\mu - \frac{\sigma}{|\xi|}, \infty), & \text{for } \xi > 0 \\ (-\infty, \infty), & \text{for } \xi = 0 \\ (-\infty, \mu + \frac{\sigma}{|\xi|}], & \text{for } \xi < 0 \end{array} \right\}.$$

Similarly to the specialized case of GEV, the ultimate limits can be found $x^* = \mu - \sigma/|\xi|$ if $\xi > 0$ and $x^* = \mu + \sigma/|\xi|$ if $\xi < 0$. I implemented and evaluated the fit of various location parameters including an evaluated Gompertz curve and a constant value.

3.2 Fitting GEV/GEVr Location Parameter

3.2.1 Fitting a Constant as the Location Parameter

A constant k for the location parameter μ in the GEV distribution is the simplest usage:

$$f(x) = \frac{1}{\sigma}\left(1 + \xi\frac{x-k}{\sigma}\right)^{-\frac{1}{\xi}-1} e^{-\left(1+\xi\frac{x-k}{\sigma}\right)^{-\frac{1}{\xi}}}. \tag{3}$$

And similarly, a constant implemented into the GEVr distribution's probability density function:

$$f_r(x_1, x_2, ..., x_r|k, \sigma, \xi) = \sigma^{-r} \exp\left\{-(1+\xi z_r)^{-\frac{1}{\xi}} - \left(\frac{1}{\xi}+1\right)\sum_{j=1}^{r} \log(1+\xi z_j)\right\}, \tag{4}$$

where $x_1 > \cdots > x_r$, $z_j = (x_j - k)/\sigma$, and $1 + \xi z_j > 0$ for $j = 1, \ldots, r$. A resulting upper limit $x^* = k + \sigma/|\xi|$ or lower limit $x^* = k - \sigma/|\xi|$ can then be found.

3.2.2 Fitting a Gompertz Curve as the Location Parameter

A curve can add more information about the location of the curve especially if the data is non-stationary. One reason the Gompertz curve is a meaningful implementation is because of its asymptotic behavior with its tails. Due to the asymptotic feature of the Gompertz curve, I can fit a Gompertz curve $f(t)$ as the location parameter μ of the GEV distribution thus the findings do not depend on time. A quick proof shows this to be true: $\lim_{t\to\infty} f(t) = \lim_{t\to\infty} ae^{-be^{(-ct)}} + z = ae^{-be^{(-\infty)}} + z = ae^{0} + z = a + z$. This is unlike many time series approaches like ARIMA modeling, that use time to forecast future results. This methodology allows finding an ultimate upper limit that does not have to happen at a certain time. The Gompertz approach uses the Gompertz function defined as:

$$f(t) = ae^{-be^{(-ct)}} + z. \tag{5}$$

The Gompertz curve is a function of time t. a is the asymptotic parameter (technically $a + z$ is the asymptote). b describes where the curve is placed on the x axis. c is the growth rate. z is the intercept. Here the Gompertz curve is implemented into the GEV distribution's location parameter with the resulting probability density function:

$$f(x|(f(t)), \sigma, \xi) = \frac{1}{\sigma}\left(1 + \xi\frac{x - f(t)}{\sigma}\right)^{-\frac{1}{\xi}-1} e^{-\left(1+\xi\frac{x-f(t)}{\sigma}\right)^{-\frac{1}{\xi}}}. \tag{6}$$

Similarly, the Gompertz curve implemented into the GEVr distribution is as follows:

$$f_r(x_1, x_2, ..., x_r|f(t), \sigma, \xi) = \sigma^{-r} \exp\left\{-(1 + \xi z_r)^{-\frac{1}{\xi}} - \left(\frac{1}{\xi} + 1\right)\sum_{j=1}^{r} \log(1 + \xi z_j)\right\}, \tag{7}$$

where $x_1 > \cdots > x_r$, $z_j = (x_j - f(t))/\sigma$, and $1 + \xi z_j > 0$ for $j = 1, \ldots, r$. An upper limit $x^* = (z + a) + \sigma/|\xi|$ or lower limit $x^* = (z + a) - \sigma/|\xi|$ can then be found.

3.3 Optimizing GEVr Parameters

As with any model tuning, it is important to optimize parameters in your functions. The initial values fed into the optimization algorithm were found using nonlinear least squares estimation based off of a monotone shape-restricted splines (Wang and Yan, 2017) matrix fitting of the data or using the probability weighted moments. The optimization technique applied here is the Nelder-Mead algorithm. This algorithm is a good choice because it is a nonlinear approach that does not require any derivative calculations that other methods may require, although it may be slightly more computationally expensive. Parameters for the GEV distribution are fitted to the data through the optimization method Nelder-Mead using maximum log-likelihood estimation (MLE) criteria. An important note is that the GEV distribution does not restrict the limit to be 0% to 100% for shooting percentage which is the practical absolute limit in real life.

The first technique to rectify this predicament was to apply bounded transformations such as Fisher's Z transformation and the logistic link function to the data and then continue with the methodology; however, this then changed the behavior of the underlying distribution and parameters. For example, the GEV shape parameter ξ in the NBA free throw data was negative before transformation and positive after transformation. This is problematic, so an alternative technique was tried instead which is similar to how the scale parameter σ is restricted to be positive in the optimization. Within the maximum log-likelihood calculation in the optimization algorithm, if a scale parameter at any point has a negative value then the log-likelihood function for that step's given parameters returns negative infinity. Because the optimization algorithm is finding ideal parameters based on maximizing the returned value of the log-likelihood function, a set of parameters with negative infinity as the log likelihood would surely not be chosen. So similarly, if $x^* < 0$ when $\xi > 0$ or if $x^* > 1$ when $\xi < 0$, then the log-likelihood function for that step's given parameters returns negative infinity. This ensures only estimated ultimate limits between 0% and 100%. Once the parameters are optimized (the Nelder-Mead algorithm converges) and the limit is found, an approx. parametric bootstrapped confidence interval is estimated.

3.4 95% Parametric Bootstrapped Confidence Interval

To get a sense of the range of what the true ultimate limit could actually be, a 95% parametric bootstrapped confidence interval is calculated for each ultimate limit found. The algorithm used is laid out below:

1. 1000 bootstrapped randomly generated GEVr datasets are simulated using the optimized parameters found for the particular limit.
2. For each simulated dataset, the optimization method outlined in Section 3.3 is used to estimate each dataset's own ideal GEVr parameters.
3. For each of these simulated datasets with their own respective optimized parameters, an ultimate limit is calculated.
4. The goodness of fit is evaluated using either the Kolmogorov-Smirnov testing procedure or the Score testing procedure.
5. The simulated datasets' limits are then sorted from least to greatest and the 25th and 975th limits are the bounds of the 95% confidence interval.

3.5 Goodness of Fit Testing

3.5.1 GEV GOF Testing

To evaluate goodness of fit for the GEV distribution, I use the Kolmogorov-Smirnov test on my data after I have subtracted the location from it. I cannot use the pvalue from the output of this test however because it relies on the mean from the data so instead a parametric bootstrap approach

each of the 1000 simulated datasets, again after subtracting the location from each data point, as described in Section 3.4. A simulated approximate p value can be found by finding the number of simulated datasets' Kolmogorov-Smirnov test statistics that are greater than the original dataset's Kolmogorov-Smirnov test statistic and divide that by 1000. If the approximated pvalue is < 0.05, then it means that the distribution is not a good fit to the data. A histogram of these simulated datasets' Kolmogorov-Smirnov test statistics with the original dataset's Kolmogorov-Smirnov test statistic overlaid on top of can be shown to help illustrate this approximated pvalue.

Table 1: Estimated Limits from Various Methods.

Data	r	Method	Lower CR	Upper CR	Upper/Lower	Limit	95% CI
FT	1	Constant	0.843	0.981	Upper	0.992	(0.952,1.000)
FT	1	Gompertz	0.843	0.981	Upper	0.996	(0.969,1.000)
FT	3	Constant	0.843	0.981	Upper	0.997	(0.960,1.000)
FT	3	Gompertz	0.843	0.981	Upper	0.995	(0.975,1.000)
3pt	1	Constant	0.441	0.536	Lower	0.407	(0.045,0.427)
2pt	1	Gompertz	0.447	0.727	Lower	0.429	(0.095,0.544)

3.5.2 GEVr GOF Testing

When there is a matrix with multiple columns of data for each observed year, then the Kolmogorov-Smirnov testing procedure outlined above cannot be used because the Kolmogorov-Smirnov test takes a single vector of observed data as its argument. So as an alternative, a Parametric Bootstrapped Score test is applied. This score test uses the score function and Fisher's Information matrix to create a score statistic and this is simulated 1000 times to return an approximate p value similar to the Kolmogorov-Smirnov bootstrapping procedure.

4 Results

Table 1 is a table of results for all of the ultimate limits. The columns of the table represent the dataset used, the r value in the GEVr distribution, method used (which location parameter for GEV dist.), the current lowest league leading record, the current highest league, the type of estimated ultimate limit found (upper or lower), the estimated ultimate limit found, and the corresponding 95% confidence interval for that estimated ultimate limit. The type of estimated ultimate limit found (upper or lower) is found from the GEV shape parameter, i.e., if the shape parameter is positive, only an ultimate lower limit can be found and if the shape parameter is negative, only an ultimate upper limit can be found. An ultimate upper limit can be interpreted as the highest an NBA player can conceivably shoot for a season. And conversely, an ultimate lower limit can be interpreted as the lowest an NBA player can conceivably shoot for a season and still lead the league in shooting percentage. A lower limit is NOT the lowest an NBA player can shoot in a season.

4.1 Free Throw

4.1.1 Constant Location (GEV)

The limit is calculated with Table 2's optimized parameter values. Figure 2 shows the GEV (Constant Location Parameter) optimized parameters fitted to the NBA Free Throw data. The ultimate upper limit is $x^* = \mu + \sigma/|\xi| = 0.915 + 0.018/| - 0.236| = 0.992$. So according to this method, the highest an NBA player can conceivably shoot free throws for a season is 99.2%. The 95% approx. parametric bootstrap confidence interval was calculated to be $(0.952, 1.000)$. The corresponding approx. pvalue = 0.415 from the Kolmogorov-Smirnov tests. $0.415 > 0.05$, so therefore the goodness of fit is satisfied for this GEV distribution with constant location parameter. Also a histogram

Table 2: Optimized Parameters for NBA Free Throw Percentage (Constant GEV Location)

GEV Parameters	Optimized Parameter Values	95% Bootstrapped Confidence Interval
Location	0.915	(0.881,0.966)
Scale	0.018	(0.015,0.023)
Shape	-0.236	(-0.195,-0.601)

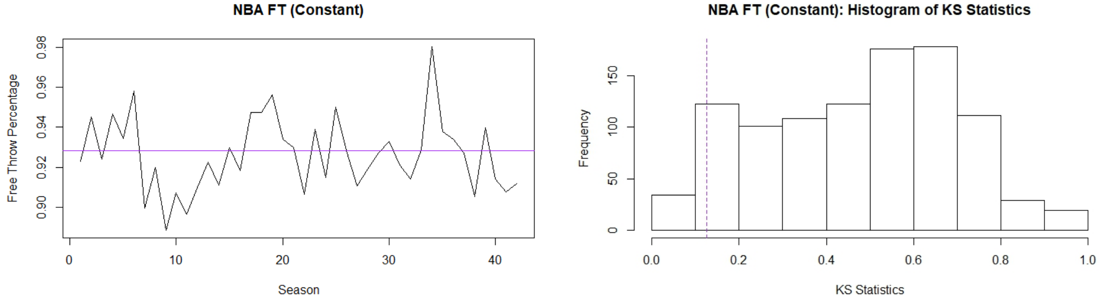

Figure 2: Constant GEV MLE fitting to Free Throw Data

displaying the Kolmogorov-Smirnov bootstrapped test statistics is shown in Figure 2 with the main Kolmogorov-Smirnov test statistic of the derived parameters for the shooting percentage data overlaid on the top layer. Because the main Kolmogorov-Smirnov test statistic is not very extreme on the upper tail this indicates a good fit for the data.

4.1.2 Gompertz Location (GEV)

The limit is calculated with Table 3's optimized parameter values. Figure 3 shows the GEV (Gompertz Location Parameter) optimized parameters fitted to the NBA Free Throw data. The ultimate upper limit is $x^* = \mu + \sigma/|\xi| = (0.867 + 0.061) + 0.024/| - 0.353| = 0.996$. So according to this method, the highest an NBA player can conceivably shoot free throws for a season is 99.6%. The 95% approx. parametric bootstrap confidence interval was calculated to be $(0.969, 1.000)$. The corresponding approx. pvalue = 0.415 from the Kolmogorov-Smirnov tests. $0.415 > 0.05$, so therefore the goodness of fit is satisfied for this GEV distribution with Gompertz location parameter.

Table 3: Optimized Parameters for NBA Free Throw Percentage (Gompertz GEV Location)

GEV Parameters	Optimized Parameter Values	95% Bootstrapped Confidence Interval
z	0.867	(0.851,0.882)
a	0.061	(0.045,0.078)
b	3.332	(3.253,3.811)
c	0.068	(0.044,0.116)
Scale	0.024	(0.019,0.028)
Shape	-0.353	(-0.265,-0.594)

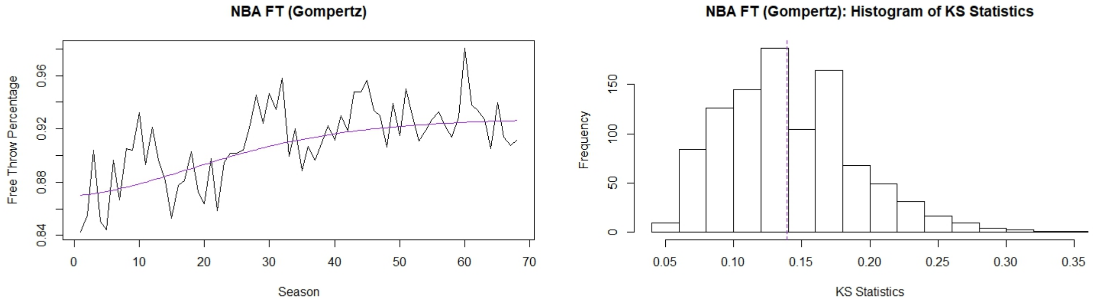

Figure 3: Gompertz GEV MLE fitting to Free Throw Data

Table 4: Optimized Parameters for NBA Free Throw Percentage (Constant GEVr($r = 3$) Location)

GEVr Parameters	Optimized Parameter Values	95% Bootstrapped Confidence Interval
Location	0.919	(0.914,0.924)
Scale	0.019	(0.016,0.021)
Shape	-0.240	(-0.124,-0.356)

Also a histogram displaying the Kolmogorov-Smirnov bootstrapped test statistics is shown in Figure 3 with the main Kolmogorov-Smirnov test statistic of the derived parameters for the shooting percentage data overlaid on the top layer. Because the main Kolmogorov-Smirnov test statistic is not very extreme on the upper tail this indicates a good fit for the data.

4.1.3 Constant GEVr(r=3)

The limit is calculated with Table 4's optimized parameter values. Figure 4 shows the GEVr($r = 3$) (Constant Location Parameter) optimized parameters fitted to the NBA Free Throw data. The ultimate upper limit is $x^* = \mu + \sigma/|\xi| = (0.919) + 0.019/| - 0.240| = 0.997$. So according to this method, the highest an NBA player can conceivably shoot free throws for a season is 99.7%. The 95% approx. parametric bootstrap confidence interval was calculated to be $(0.975, 1.000)$. The corresponding approx. pvalue = 0.831 from the bootstrap Score Test. $0.831 > 0.05$, so therefore the goodness of fit is satisfied for this GEVr distribution with constant location parameter. The confidence interval is narrower when analyzing the top 3 $(0.960, 1.000)$ than it is when only analyzing the top 1 $(0.952, 1.000)$ shooting percentage per year. This is a major advantage to using more data in your estimation. It is important to consider the value of r that you select however. If you use too many top performers your goodness of fit could be poor so I selected the top 3 ($r = 3$) as a modest choice. In some future work, I could implement some sequential testing to find the optimal r to use.

4.1.4 Gompertz GEVr(r=3)

The limit is calculated with Table 5's optimized parameter values. Figure 5 shows the GEVr($r = 3$) (Gompertz Location Parameter) optimized parameters fitted to the NBA Free Throw data. The ultimate upper limit is $x^* = \mu + \sigma/|\xi| = (0.872 + 0.061) + 0.023/| - 0.361| = 0.995$. So according to this method, the highest an NBA player can conceivably shoot free throws for a season is 99.5%.

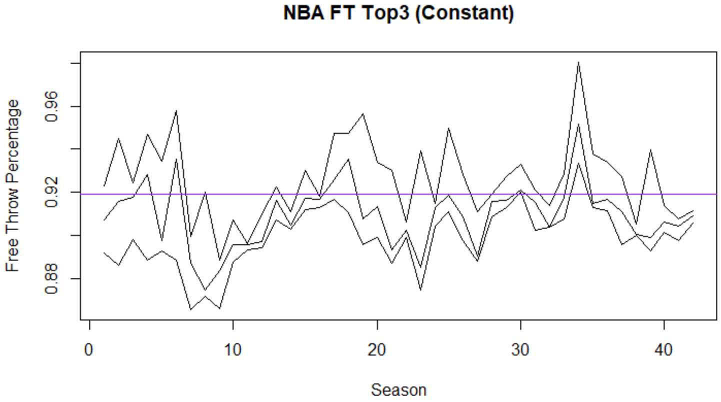

Figure 4: Constant GEVr($r = 3$) MLE fitting to Free Throw Data

Table 5: Optimized Parameters for NBA Free Throw Percentage (Gompertz GEVr($r = 3$) Location)

GEVr Parameters	Optimized Parameter Values	95% Bootstrapped Confidence Interval
z	0.8720	(0.861,0.881)
a	0.0610	(0.049,0.073)
b	5.5310	(5.470,6.154)
c	0.0754	(0.060,0.102)
Scale	0.0230	(0.020,0.024)
Shape	-0.3610	(-0.298,-0.497)

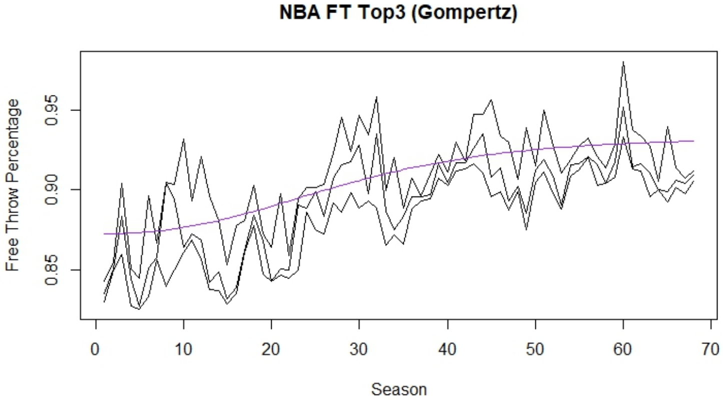

Figure 5: Gompertz GEVr($r = 3$) MLE fitting to Free Throw Data

Table 6: Optimized Parameters for NBA 3 Point Percentage (Constant GEV Location)

GEV Parameters	Optimized Parameter Values	95% Bootstrapped Confidence Interval
Location	0.467	(0.460,0.474)
Scale	0.019	(0.013,0.024)
Shape	0.313	(0.063,0.564)

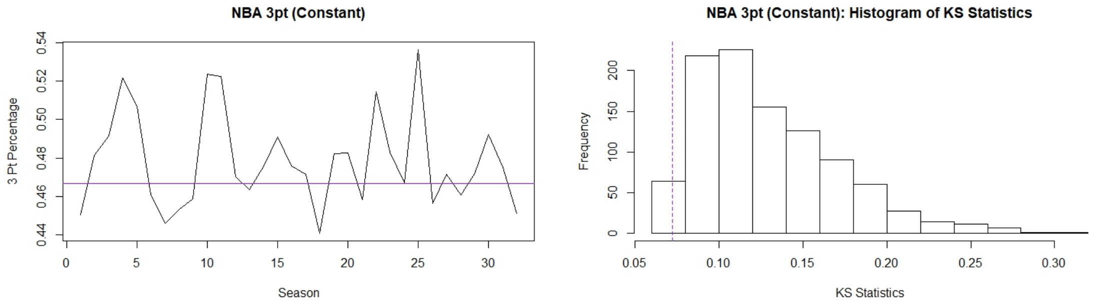

Figure 6: Constant GEV MLE fitting to 3 Point Data

The 95% approx. parametric bootstrap confidence interval was calculated to be $(0.975, 1.000)$. Once again, the confidence interval is narrower when analyzing the top 3 $(0.975, 1.000)$ than it is when only analyzing the top 1 $(0.969, 1.000)$ shooting percentage per year. This is a major advantage to using more data in your estimation. The Parametric Bootstrap Test within the *eva* (Bader, 2016) R package can only be performed with a constant location parameter for the GEVr fitting and in this example I used a Gompertz curve. In some future work, I can adapt this methodology for non-constant location parameters. So in this paper, I have not verified the goodness of fit for these GEVr parameters when r=3.

Overall for all the methods applied to finding an ultimate upper limit, the limit was above 99.2% which would exceed the current NBA single season record for free throw percentage. However, 100% was within the confidence interval for each estimation as well. So it may truly be possible for one player to make all of their shots although I estimate that the highest free throw percentage achievable is slightly below this perfect 100%.

4.2 3 Point Field Goal Percentage

4.2.1 Constant Location GEV

The limit is calculated with Table 6's optimized parameter values. Figure 6 shows the GEV (Constant Location Parameter) optimized parameters fitted to the NBA 3 Point data. The ultimate lower limit is $x^* = \mu + \sigma/|\xi| = 0.467 - 0.019/|0.313| = 0.407$. So according to this method, the lowest percentage an NBA player can conceivably shoot for 3 pointers while still leading the league is 40.7%. The 95% approx. parametric bootstrap confidence interval was calculated to be $(0.045, 0.427)$. The corresponding approx. pvalue = 0.989 from the Kolmogorov-Smirnov tests. $0.989 > 0.05$, so therefore the goodness of fit is satisfied for this GEV distribution with constant location parameter. Also a histogram displaying the Kolmogorov-Smirnov bootstrapped test statistics is shown in Figure

Table 7: Optimized Parameters for NBA 2 Point Percentage (Gompertz GEV Location)

GEV Parameters	Optimized Parameter Values	95% Bootstrapped Confidence Interval
z	0.439	(0.403,0.468)
a	0.165	(0.135,0.202)
b	4.136	(2.095,10.101)
c	0.118	(0.086,0.164)
Scale	0.027	(0.020,0.032)
Shape	0.152	(0.037,0.390)

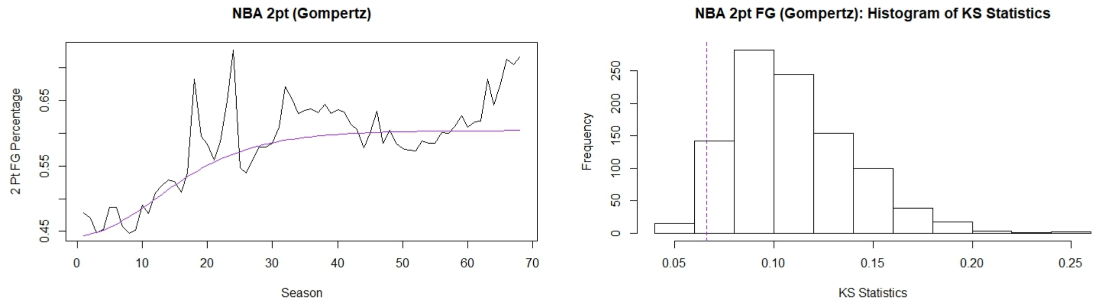

Figure 7: Gompertz GEV MLE fitting 2 Point Data

6 with the main Kolmogorov-Smirnov test statistic of the derived parameters for the shooting percentage data overlaid on the top layer. Because the main Kolmogorov-Smirnov test statistic is not very extreme on the upper tail this indicates a good fit for the data.

4.3 2 Point Field Goal Percentage

4.3.1 Gompertz Location GEV

The limit is calculated with Table 7's optimized parameter values. Figure 7 shows the GEV (Gompertz Location Parameter) optimized parameters fitted to the NBA 2 Point data. The ultimate lower limit is $x^* = \mu + \sigma/|\xi| = (0.439 + 0.165) - 0.027/|0.152| = 0.429$. So according to this method, the lowest percentage an NBA player can conceivably shoot for 2 pointers while still leading the league is 42.9%. The 95% approx. parametric bootstrap confidence interval was calculated to be $(0.095, 0.544)$. The corresponding approx. pvalue = 0.961 from the Kolmogorov-Smirnov tests. $0.961 > 0.05$, so therefore the goodness of fit is satisfied for this GEV distribution with Gompertz location parameter. Also a histogram displaying the Kolmogorov-Smirnov bootstrapped test statistics is shown in Figure 7 with the main Kolmogorov-Smirnov test statistic of the derived parameters for the shooting percentage data overlaid on the top layer. Because the main Kolmogorov-Smirnov test statistic is not very extreme on the upper tail this indicates a good fit for the data.

5 Discussion

The field of advanced analytics in basketball statistics is growing rapidly with many new ways to assist coaches and general managers in evaluating players. In this paper I have estimated some upper and lower limits to lead the league in free throw, 2 point field goal, and 3 point field goal percentage. For an NBA team, having the most efficient shooters on the court can help maximize wins, similar to the Money-ball Oakland Athletics from baseball. So the question of how to estimate how efficient shooters may become is important. Another important point to keep in mind is that different positions have different shooting abilities. Typically, guards are able to shoot better than forwards who are able to shoot better than centers. A breakdown of ultimate limits to shooting percentages by position would help at higher level than the limits calculated here. In this paper, only 3 main shooting percentages are analyzed. Further research on other continuous basketball statistics, for example usage percentage, assist percentage, etc., following a similar methodology could be very insightful.

In this methodology, the GEVr distribution was used to estimate the limits, and bootstrapped confidence intervals were provided as well. The choice of r, was a modest selection of $r = 1$ and $r = 3$. A future work to explore would be selecting an optimal r value through a more logically and data-driven method. This can be done by using any of the sequential testing procedures outlined in Bader et al. (2017). Sequentially for each r value a goodness of fit test can be performed and the resulting p value will be reported. The tests continue until the ideal r value which predicts the narrowest confidence interval while having an adequate goodness of fit is found. Once this optimal r value is found, following the methodology described in this paper could lead to an even smaller confidence interval for the ultimate limit while still preserving goodness of fit. The R package: *eva* (Bader, 2016) contains sequential testing with three goodness of fit tests: an entropy difference test, a parametric bootstrap score test, and a multiplier score test which is a fast weighted alternative to the parametric bootstrap score test. Another limitation in my work was the lack of assessment of goodness of fit for the Gompertz location parameter for the GEVr ($r = 3$) distribution. The current functions written can only evaluate the goodness of fit if this location parameter is a constant. Therefore future work to be done includes adapting the source code of these goodness of fit functions to allow for non-constant location parameters. I am unaware of any current packages that allow for such functionality of selecting an ideal r through goodness of fit testing for the GEVr distribution with a non-constant location parameter such as a Gompertz curve so this would be very helpful research to the community.

References

Bader, B. (2016), *eva: Extreme Value Analysis with Goodness-of-Fit Testing*, R package version 0.2.4.

Bader, B., Yan, J., and Zhang, X. (2017), “Automated selection of r for the r largest order statistics approach with adjustment for sequential testing,” *Statistics and Computing*, 27, 1435–1451.

Chiou, S., Kang, S., and Yan, J. (2015), *Extreme Value Modeling and Risk Analysis: Methods and Applications*, CRC Press.

Einmahl, J. H. J. and Magnus, J. R. (2008), “Records in Athletics through Extreme-Value Theory,” *Journal of the American Statistical Association*, 103, 1382–1391.

Einmahl, J. H. J. and Smeets, S. G. W. R. (2011), “Ultimate 100-m world records through extreme-value theory,” *Statistica Neerlandica*, 65, 32–42.

LLC, S. R. (2018), “Basketball-Reference.com - Basketball Statistics and History,” Data retrieved from Basketball-Reference database.

Marcos, M. and Woodworth, P. L. (2017), “Spatiotemporal changes in extreme sea levels along the coasts of the North Atlantic and the Gulf of Mexico,” *Journal of Geophysical Research: Oceans*, 122, 7031–7048.

Wang, W. and Yan, J. (2017), *splines2: Regression Spline Functions and Classes Too*, R package version 0.2.6.

Wickham, H. and Chang, W. (2016), *ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics*, R package version 2.2.1.

Wickham, H., Francois, R., Henry, L., and Muller, K. (2017), *dplyr: A Grammar of Data Manipulation*, R package version 0.7.4.