Modeling Dunking in the NBA using the NBD and its Variants

Matthew Pearl

February 26, 2023

1 Introduction

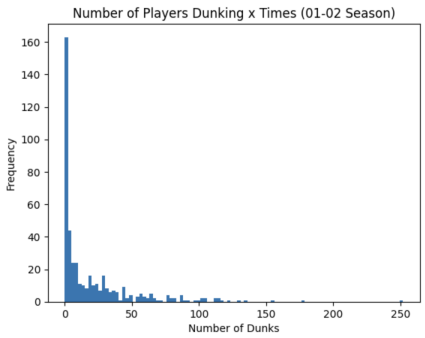
The goal of this analysis is to find a model that fits the distribution of total dunks in the NBA per season in an effective manner both quantitatively and with respect to underlying assumptions about professional basketball players. An additional objective is to observe patterns in how heterogeneity and the proportion of players who are predicted to never dunk has changed over the past 2 decades. It is hypothesized that recent seasons will lead to more heterogeneity as dunking becomes adopted by an ever-growing proportion of players over time. This, coupled with increasingly diverse play styles in the NBA, would reasonably account for this trend.

12 NBD models will be used (4 variations for each of 3 seasons) and their relevant trade-offs will be discussed. Given the nature of varying play time in the NBA alongside an assumed proportion of players who will never dunk, it is additionally hypothesized that a 0-inflated time-varying NBD will perform the best out of these models and make the most sense for this domain.

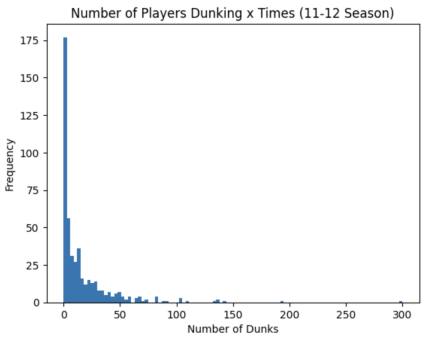
2 Data and Preprocessing

The models built in this project are for aggregate shooting data for each player in the 2001-2002, 2011-2012, and 2021-2022 NBA seasons[1]. Relevant statistics from these datasets that will be used in the analysis include the total number of minutes that a particular player played during each season as well as how many successful dunks he made.

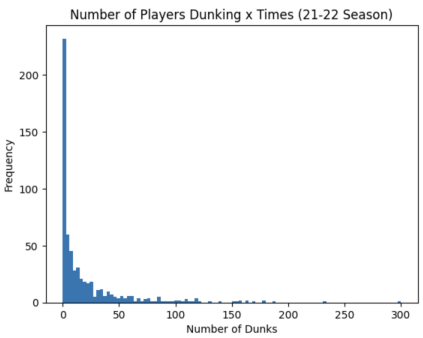
Python was used for preprocessing, with the first task being to generate count data from each of the seasons where x would represent a total number of dunks that a player could achieve over the course of a season and N_x would be the frequency for which that x was achieved. Another important detail to account for was the need to group by players, given that a player could be a part of several different teams during a season in the event that he was traded, causing them to account for multiple instances in the dataset. The following are the distributions of dunks from the aggregated data:



Dunk Distribution (2021-2022 Season)



Dunk Distribution (2011-2012 Season)



Dunk Distribution (2001-2002 Season)

[1] https://www.basketball-reference.com/

3 Model Selection

What makes the negative binomial distribution (NBD) model an appropriate choice for fitting this data is primarily the fact that dunks over the course of an NBA player's season fall within the category of count data. While there may be a theoretical limit to how many times a player could dunk in a given season, there is no finite upper bound or notion of maximum "choices", implying that models like the beta-binomial would not make sense in this setting. Furthermore, the count data is quite sparse, which is to be expected from what is a relatively rare event in the game of basketball (many players are expected to have 0 or very few dunks, and a small minority will have many). This rarity also reinforces the idea that the Poisson distribution will allow for a reasonable model of the individual level behavior - i.e., the number of times that a player dunked over the course of the season. It is also critical to note that NBA players have very different total minutes played per season, and so the time-varying NBD will prove to be useful. Moreover, the option to inflate certain probabilities corresponding to dunk totals is an additional advantage of the NBD.

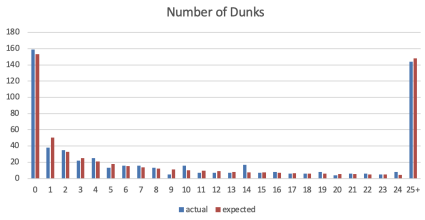
Note that for any model discussed in the following sections, a p-value of greater than 0.2 in a χ^2 goodness of fit test will be the threshold to conclude that there is significant evidence indicating a strong fit for the data. Additionally, in comparing simple and complex models in the same overarching class, a likelihood ratio test will be used and a larger model with more parameters will be said to perform significantly better than its smaller counterpart if the p-value is less than $\alpha = 0.05$. Additionally, as a rule of thumb, right-censoring will be used for χ^2 tests such that at least 80% of the expected values implied by the model are at least 5. Finally, the χ^2 test statistic for the likelihood ratio test will be $2[LL_{\text{complex}} - LL_{\text{simple}}]$ and the degrees of freedom will be $|df_{\text{complex}} - df_{\text{simple}}|$. In this section, all models will be fit using maximum likelihood estimation through minimizing the log-likelihood of the observed data.

3.1 Standard NBD on Aggregate Data

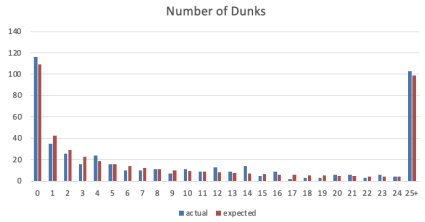
The first model will simply be a standard NBD on the aggregated data. The following table summarizes the relevant statistics:

Model Summaries			
Value	NBD ('21-22)	NBD ('11-12)	NBD ('01-02)
$\hat{r}$	0.334786	0.39913	0.349313
$\hat{\alpha}$	0.01681	0.02551	0.01755
$\max LL$	-2210.73	-1694.25	-1626.24
χ^2	29.42531	27.08439	27.75149
df	23	23	23
p-val	0.16662	0.25239	0.22532

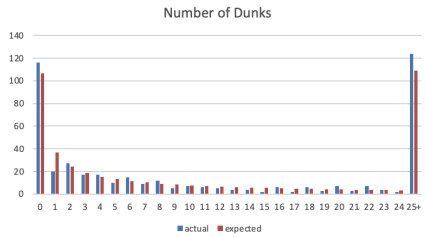
Across the board, the r values fall within the range of $0.33 - 0.4$, indicating a similar level of moderate heterogeneity in the total number of dunks that players have for the season. Out of the 3 models, only the NBDs for the 2011-2012 and 2001-2002 significantly fit the data given that their associated p-values for the χ^2 goodness of fit test are at least 0.2. The following histograms show the actual and expected counts for each season under the standard NBD on the aggregate data:



Standard NBD on Aggregate Data (2021-2022 Season)



Standard NBD on Aggregate Data (2011-2012 Season)



Standard NBD on Aggregate Data (2001-2002 Season)

3.2 Zero-Inflated NBD on Aggregate Data

In the game of basketball, there is reason to believe that certain players will never dunk (hardcore never dunkers or HNDs). 2 reasons for the occurrence of HNDs are (1) the fact that a player may be physically incapable of dunking, or (2) the plays that a player is expected to execute will never involve driving to the basket (e.g., non-starters who are pure sharp-shooters or play-makers). This leads to an extension of the previous model, where an additional parameter π will be used to create a spike at the category for 0 dunks over the course of a particular season.

The following table summarizes these more complex model adding in a row for the new π value:

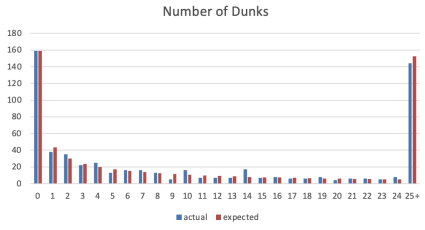
Model Summaries			
Value	0-inflated NBD ('21-22)	0-inflated NBD ('11-12)	0-inflated NBD ('01-02)
$\hat{r}$	0.41129	0.51903	0.52506
$\hat{\alpha}$	0.01888	0.02986	0.02245
$\hat{\pi}$	0.08602	0.09989	0.14904
$\max LL$	-2209.49	-1692.39	-1662.04
χ^2	26.57867	23.21796	22.76891
df	22	22	22
p-val	0.22767	0.38951	0.41488

In this case, all of the models significantly fit the data according to the p-value criterion. Moreover, the $\hat{\pi}$ values indicate that anywhere between $8-15\%$ of NBA players over the last 2 decades are projected to never dunk under any circumstances. Additionally, the r-values of between $0.4-0.53$ indicate a higher level of homogeneity in the aggregate data than the standard NBD suggested.

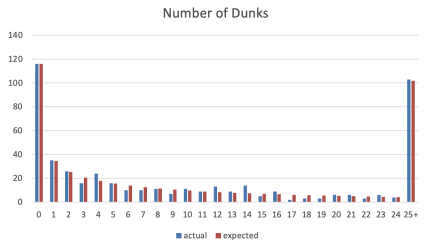
Likelihood-ratio tests will now be used to determine whether this new model with 3 parameters provides significant improvement over the standard NBD. The results are summarized in the following table:

Likelihood-Ratio Test Summaries			
Value	NBD comparison ('21-22)	NBD comparison ('11-12)	NBD comparison ('01-02)
χ^2	2.47380	3.74117	8.39085
df	1	1	1
p-val	0.11576	0.05309	0.00377

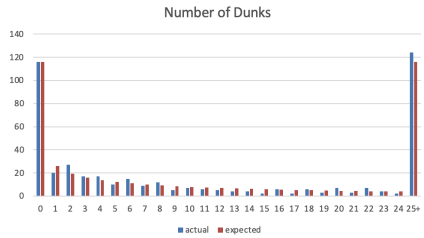
What these results show is that only the 0-inflated NBD for the 2001-2002 season significantly improves upon its corresponding standard NBD. One hypothesis to explore in further research is the idea of an "adoption curve" for dunking in the NBA over time, where less and less players in the league are considered HNDs. This will be discussed more in section 5. The following histograms show the actual and expected counts for each season under the 0-inflated NBD on the aggregate data and due to the additional parameter π, the model now fits the 0-dunks category perfectly.



0-inflated NBD on Aggregate Data (2021-2022 Season)



0-inflated NBD on Aggregate Data (2011-2012 Season)



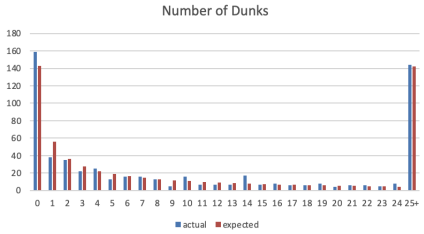
0-inflated NBD on Aggregate Data (2001-2002 Season)

3.3 Time-Varying NBD on Disaggregate Data

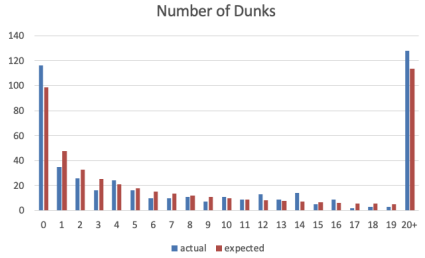
As previously mentioned, there is significant variance in the number of minutes played among NBA players. Therefore, using the season as unit time will not be sufficient and so minutes played (MP) will be used in order to capture the different levels of opportunity that a particular player had to dunk throughout the course of a season. That is, the MP feature of the data will serve as the t value in the Poisson distribution where non-unit time is taken into account - i.e., $P(X(t) = x)$ as opposed to the prior use of $P(X = x)$. The results for the time-varying models are summarized below:

Model Summaries			
Value	Time-Varying NBD ('21-22)	Time-Varying NBD ('11-12)	Time-Varying NBD ('01-02)
$\hat{r}$	0.62692	0.61622	0.55556
$\hat{\alpha}$	34.32287	41.84764	38.89074
$\max LL$	-2029.42	-1604.32	-1529.83
χ^2	34.08490	34.24770	35.03133
df	23	18	18
p-val	0.06394	0.01174	0.00937

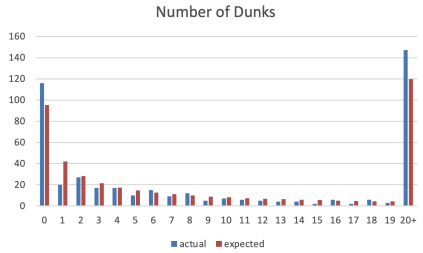
Given the low p-values, all of the time-varying NBD models fit the data relatively poorly. Assuming a significant fit, the r values yet again represent more homogeneity as compared to those suggested by models on the aggregate data. The following histograms show the actual and expected counts for each season under the standard time-varying NBD on the disaggregate data:



Time-Varying NBD on Disaggregate Data (2021-2022 Season)



Time-Varying NBD on Disaggregate Data (2011-2012 Season)



Time-Varying NBD on Disaggregate Data (2001-2002 Season)

3.4 Zero-Inflated Time-Varying NBD on Disaggregate Data

For the same reason described in the zero-inflated NBD on aggregate data, a spike at 0 will be added to the time-varying NBD on disaggregate data. The results are summarized below:

Model Summaries			
Value	0-inflated Time-Varying NBD ('21-22)	0-inflated Time-Varying NBD ('11-12)	0-inflated Time-Varying NBD ('01-02)
$\hat{r}$	0.71630	0.94572	1.0076
$\hat{\alpha}$	37.61510	56.76480	59.75989
$\hat{\pi}$	0.04018	0.11681	0.15322
$\max LL$	-2027.97	-1595.25	-1513.27
χ^2	29.89367	24.02467	18.54776
df	22	22	17
p-val	0.12107	0.34594	0.35510

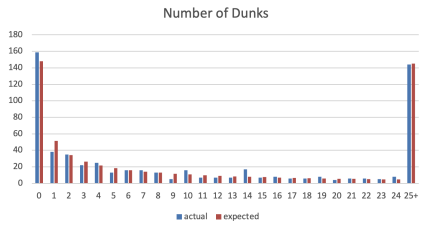
The models for the 2011-2012 and 2001-2002 seasons fit the data quite well and all of the models have a non-zero π value, predicting that anywhere from $4 - 16\%$ of players are HNDs. Again, another interesting research direction could be to analyze the change in this π value over a larger time frame. Interestingly, the model for 2001-2002 indicates a predicted r value of greater than 1 and all of the models suggest a high level of homogeneity, contrasting

some of the conclusions made from previous models.

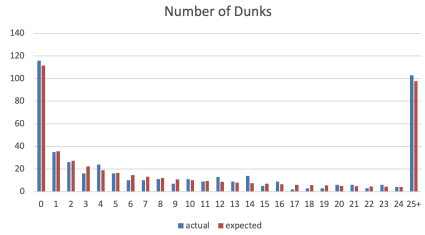
As in section 3.2, a likelihood-ratio test will be used to determine whether these improvements upon the standard time-varying NBD are significant:

Likelihood-Ratio Test Summaries			
Value	Time-Varying NBD comparison ('21-22)	Time-Varying NBD comparison ('11-12)	Time-Varying NBD comparison ('01-02)
χ^2	2.90854	18.13468	35.11519
df	1	4	1
p-val	0.08811	0.00116	≈ 0

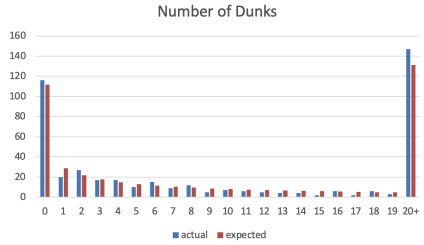
This test indicates that only the models for 2011-2012 and 2001-2002 make significant improvements over the standard NBD. Out of all of the model types, the 0-inflated time-varying NBD will be the model of choice, as it fits the data quite well, but more importantly tells the most reasonable story about NBA players - i.e., that they have very different levels of playing time during the regular season and that there is expected to be some non-zero portion of HNDs. The following histograms show the actual and expected counts for each season under the 0-inflated time-varying NBD on the disaggregate data:



0-inflated Time-Varying NBD on Disaggregate Data (2021-2022 Season)



0-inflated Time-Varying NBD on Disaggregate Data (2011-2012 Season)



0-inflated Time-Varying NBD on Disaggregate Data (2001-2002 Season)

4 Model Robustness

Robustness will be tested using the aggregated (non time-varying data), where it is desirable to obtain similar r and α parameters using methodologies other than maximum likelihood estimation. Here, the "Method of Moments" (MoM) and "Means and Zeros" (M&Z) approaches will be used. The parameters will first be estimated using the MoM, where

$$\hat{\alpha} = \frac{\bar{x}}{s^2 - \bar{x}} \text{ and } \hat{r} = \hat{\alpha}\bar{x}$$

For the standard NBD on the aggregate data, the MoM predicted the values for r and α quite poorly as demonstrated in the following table. This is likely a result of the high mean due to outliers.

Likelihood-Ratio Test Summaries			
Value	MoM ('21-22)	MoM ('11-12)	MoM ('01-02)
$\bar{x}$	46.13411	39.84277	73.45
s^2	1784.15	1166.58	3807.89
$\hat{r}$	1.22459	1.40888	1.44464
$\hat{\alpha}$	0.02654	0.03536	0.01967
χ^2	3769.53	3467.01	35538.17
df	23	23	23
p-val	≈ 0	≈ 0	≈ 0

While they did not fit the data well, the M&Z approach created much more reasonable predictions for r and α that more closely resembled the parameters derived from maximum likelihood estimation, indicating some robustness

for the model. In particular, r and α will be estimated using the system

$$P_0 = (\frac{\hat{\alpha}}{\hat{\alpha}+1})^{\hat{\alpha}\bar{x}} \text{ and } \hat{r} = \hat{\alpha}\bar{x}$$

Likelihood-Ratio Test Summaries			
Value	MoM ('21-22)	MoM ('11-12)	MoM ('01-02)
$\bar{x}$	46.13411	39.84277	73.45
P_0	0.26325	0.24319	0.26364
$\hat{r}$	0.25697	0.28607	0.23137
$\hat{\alpha}$	0.00557	0.00718	0.00315
χ^2	84.69301	82.41488	71.41444
df	23	23	23
p-val	≈ 0	≈ 0	≈ 0

Clearly, these estimated r values are much more consistent with the levels of heterogeneity determined using MLE. While the aggregate model may not be robust, additional techniques should be used for analyzing the robustness of the time-varying models.

5 Limitations and Future Research Objectives

The primary limitation of the data was the unavailability of older statistics on dunking in the NBA. It would have been interesting to see the changes in heterogeneity and the proportion of HNDs over a longer time-frame, i.e. from when the first players started dunking in the 1960s up until today. It may be worthwhile to consider all of the seasons between 2001 and 2023 and use regression to understand how these parameters have changed over time. The limited data on this subject ultimately made it difficult to gauge the overall evolution of dunking in the NBA and only allows for speculation that would need to be more rigorously tested.

An additional area of interest is to explore the effects of covariates on an NBA player's propensity to dunk. There are many ways that the data can be grouped to understand how a player's position, height, team, whether they are an all-star, etc. might affect this propensity and it is almost certain that different subsets of players based on these covariates would lead to very different parameters. For instance, it may be expected that there is a high level of homogeneity among the dunking numbers for starting centers, whereas point guards will have a lot more heterogeneity given that some with an athletic playing style will have a higher propensity to dunk and others may fall into the category of HNDs.

Finally, additional research on the robustness of the time-varying models using more advanced approaches for parameter estimation should be conducted.

6 Conclusion

As hypothesized, the 0-inflated time-varying NBD performed the best for the disaggregate data. In addition to the nature of the data, this particular model tells the best story since players have varying play time and some never dunk due to their inability to do so or because their play style will never call for it.

The consideration of HNDs in the final model made significant improvements upon the standard time-varying NBD in the cases of the 2011-2012 and 2001-2002 seasons and the final r values indicate more homogeneity than originally expected, however they do follow the hypothesized pattern of increasing heterogeneity in dunk frequencies over time. Moreover, the decreasing π values over the last 2 decades could be indicative of a trend that the proportion of players dunking in the NBA has steadily increased and may continue to in future seasons.