NBD Modeling of MLB Triples

Considering the game's rich statistical data, it's no surprise that baseball was among the earliest sports to embrace a move toward analytics-based decision-making. With numerous player-specific metrics, a wide range of outcomes, and detailed player recordings, MLB statistics provide an ideal basis for investigating unusual occurrences. Of course, some things almost never happen, like an inside-the-park walk-off home run or a pitcher-hit grand slam, but triples live in an interesting place between the mundane and unusual. For instance, in this 2022 season, 305 out of 692 players with at least one batting attempt hit a triple. Yet, the league leader only recorded 9. This project aims to identify relevant segmentations, better understand the nature of MLB players' propensities to hit these hyper-dispersed triples, and develop models that best capture such complexities.

Often referred to as the most exciting play in baseball, a triple is defined by a batter hitting the ball into play and then rounding first and second base, touching each sequentially before finally reaching third base prior to being tagged by a defender possessing the ball. If a player makes it all the way to third on his lone swing of the bat, and the defense did not aid his ability to do so as a result of "an intervening error or attempt to put out another baserunner,"[1] then this goes in the books as a triple (see *Figure 1*).[2] That last part is crucial, as it attempts to remove extraneous factors of luck that don't accurately reflect the true nature of the batter's play.

In this analysis, t-adjusted, player-segmented, and spike variations of the count-based NBD model are used exclusively, despite this technically being a choice process. The number of triples

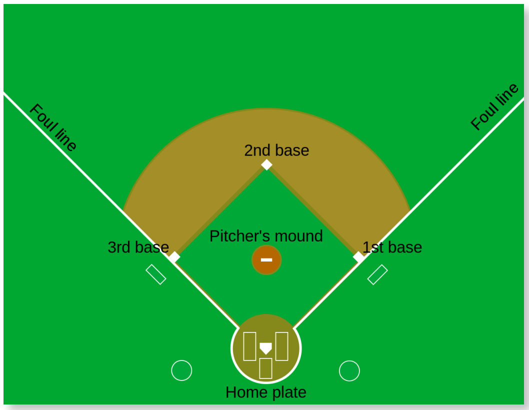
[1] "Triple (3B): Glossary." *MLB.com*, https://www.mlb.com/glossary/standard-stats/triple.

[2] "File:Baseball Diamond Simplified.svg." *Wikimedia Commons*, https://commons.wikimedia.org/wiki/File:Baseball_diamond_simplified.svg.

a player hits is limited to the number of opportunities he has at the plate. However, this upper bound is effectively unreachable. The best performer – even among those with as few as a single batting attempt (BA) – hit triples on fewer than 7% of his BAs. This one at-bat minimum serves as the cutoff for inclusion in this project's dataset, encompassing batting activity from the 2022 MLB regular season.[3]

Two segmentations were run across all of the above NBD variations: one based on players' batting hands and the other on their sprint speeds. Each of these factors into the prospect of hitting triples from a different perspective. A player's speed has an obvious impact on the amount of time required to reach third base, which is a challenging feat due to the long distance that players must run over a typically short window of time: "Because of the nature of a triple – with the batter covering three bases, or 270 feet – there is almost always a close play at third base."[4] The idea behind segmenting by speed is simple. Fast players should be able to hit more triples than slower players, as they can round the bases more quickly. A separate data source containing sprint speeds from a slightly smaller group of players (622) was merged with the original triples data, allowing for this speed-segmented analysis (see *Figure 2* for a sample of the combined data).[5] Also, it's worth noting that all speed-segmented LRT tests were run against this smaller 622-player pooled model, assuring that the same data was used in both.

Figure 1: Baseball Field Diagram

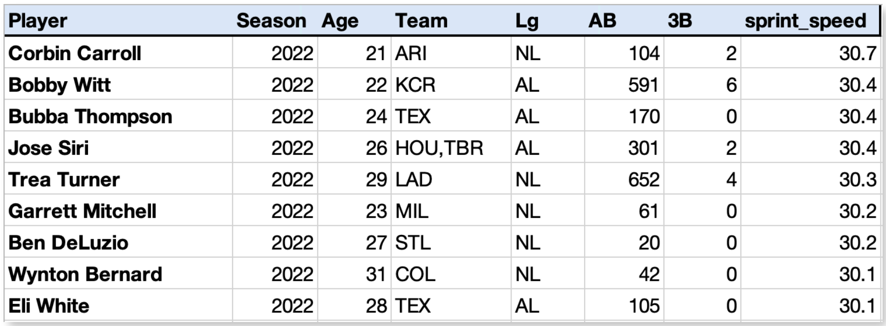


[3] "Player Batting Season & Career Stats Finder." *Stathead.com*, https://stathead.com/baseball/player-batting-season-finder.cgi.

[4] "Triple (3B): Glossary." *MLB.com*, https://www.mlb.com/glossary/standard-stats/triple.

[5] "Statcast Sprint Speed Leaderboard." *Baseballsavant.com*, https://baseballsavant.mlb.com/leaderboard/sprint_speed?min_season=2022&max_season=2022&position=all&team=&min=0.

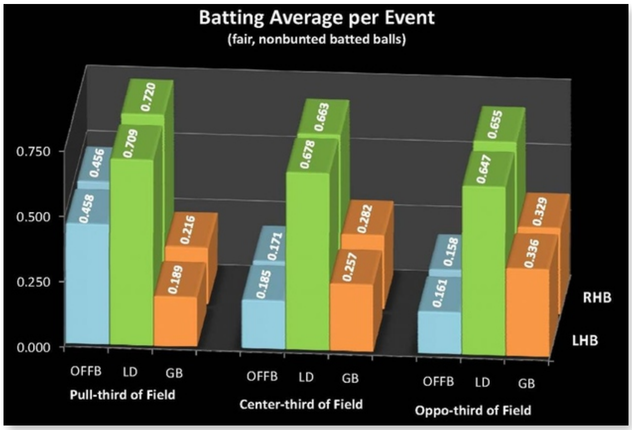
Figure 2: Sample Data



Player	Season	Age	Team	Lg	AB	3B	sprint_speed
Corbin Carroll	2022	21	ARI	NL	104	2	30.7
Bobby Witt	2022	22	KCR	AL	591	6	30.4
Bubba Thompson	2022	24	TEX	AL	170	0	30.4
Jose Siri	2022	26	HOU,TBR	AL	301	2	30.4
Trea Turner	2022	29	LAD	NL	652	4	30.3
Garrett Mitchell	2022	23	MIL	NL	61	0	30.2
Ben DeLuzio	2022	27	STL	NL	20	0	30.2
Wynton Bernard	2022	31	COL	NL	42	0	30.1
Eli White	2022	28	TEX	AL	105	0	30.1

While speed is clearly linked with one's ability to hit triples, a player's batting hand might initially seem unrelated. The significance lies primarily in its determination of a hitter's pull direction. Pulling the ball means swinging earlier, getting the bat around and through it, and sending it in the opposite direction of one's batting hand. Hitters are more successful at pulling line drives and outfield flies (the types of hits deep enough to translate into triples) than sending these balls into the opposite field (see *Figure 3*).[6] Moreover, pulled shots have more power behind them, as evidenced by a nearly 5 MPH average exit velocity increase over pushed balls into the opposite field (see *Figure 4*).[7] For right-handed hitters, who line up on the left side of the batter's box, with their left shoulder closer to the pitcher, this means swinging ahead of the ball and pulling it towards left field. The opposite is true for left-handed hitters, who pull the ball right.

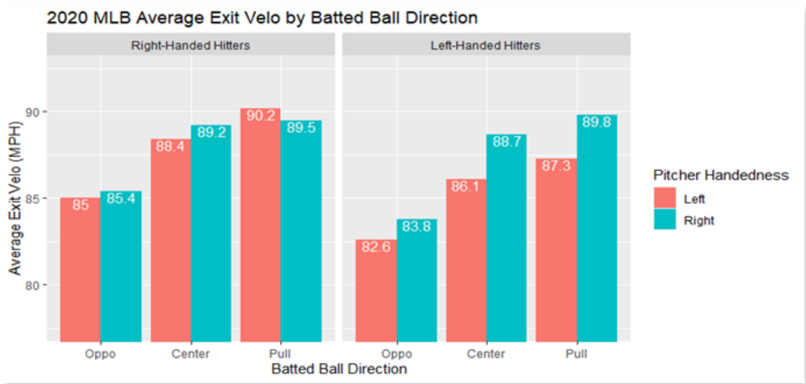
Figure 3



[6] Reillocity. "Outcomes of Flyballs, Line Drives, and Groundballs by Direction: 2013 Data from Al Parks and Minutemaid." *The Crawfish Boxes*, 9 Sept. 2013, https://www.crawfishboxes.com/2013/9/9/4677304/outcomes-of-flyballs-line-drives-and-groundballs-by-direction-2013.

[7] Published by John Moore View all posts by John Moore, et al. "Importance of Maintaining Exit Velocity to All Fields." *BaseballCloud Blog*, 16 Nov. 2020, https://baseballcloud.blog/2020/11/16/importance-of-maintaining-exit-velocity-to-all-fields/.

Figure 4



A player’s pull direction is important because it dictates the side towards which he naturally hits the ball well and with good power. Since third base is on the left side of the infield, a player gives himself more time to reach this destination by expanding the distance between the hit ball and this base, where he must finish his run before being tagged out. A ball hit into right field requires a much longer throw (or series of throws) than one hammered into left or center field. In layman’s terms, righties tend to hit the ball closer to third base, while lefties tend to hit the ball farther from third base. A farther-hit ball from where a runner is heading means he has more time to get there.

Finally, regarding batting hands, it’s worth addressing switch-hitters, who comprise just over 10% of batters in this dataset. These rare players tend to hit with the hand opposite of the pitcher’s that’s delivering the ball (meaning left when facing righties and vice versa) as it gives them what is referred to in baseball as a “platoon advantage.”[8] Since 72% of pitches come from right-handed tossers, switch hitters tend to bat left-handed.[9] For this reason and due to their small sample size, which was challenging to segment as its own subgroup, they were included with the lefties.

[8] “Platoon.” *Platoon - BR Bullpen*, https://www.baseball-reference.com/bullpen/Platoon.

[9] Birnbaum, Guy Molyneux and Phil. “The Southpaw Advantage.” *FanGraphs Baseball*, 8 Sept. 2020, https://blogs.fangraphs.com/the-southpaw-advantage/.

Despite the introduction of errors in the scorebooks to cancel out unearned hits, such as triples that result from errant defensive throws, there is still an unavoidable element of randomness involved in batting statistics. For example, if a ball takes a wild, unexpected bounce off a wall, flying past an outfielder who would otherwise be playing its position well, this would not be an error. Similarly, there have been instances where a high fly ball is lost in the sun or is hit between two outfielders, such that no play is made on the ball and no error is assessed. How is this? First, errors don't include unlucky (or lucky for the batter) bounces. Second, they are somewhat arbitrary, ultimately coming down to the discretion of the official scorer, who has the final say.[10] In short, while triples are primarily the result of skill, there is also an inherent element of chance.

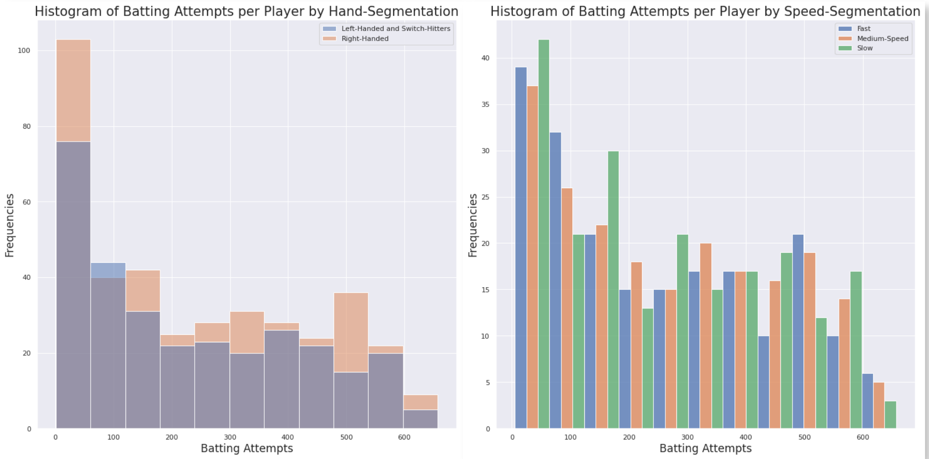
With this in mind, a running question throughout this triples analysis is whether or not a spike is warranted at zero. Are there players who would never hit a triple if given infinite opportunities under the assumption of stationarity? Some might argue that the answer is yes, citing slower players and pitchers whose focus isn't on hitting. In response to the latter point, it's important to note that the designated hitter rule went into effect this last season, meaning pitchers were no longer required to take turns batting. Nevertheless, there were rare scenarios where pitchers still stepped into the batter's box. For example, four pitchers not named Shohei Ohtani (an incredibly rare two-way pitcher/designated-hitter all-star) hit in 2022, seeing a paltry combined total of 5 at-bats. These players failed to record triples over this minute sample size, with all falling short of reaching base entirely. Can we classify a small subset of players for whom such little data exists as hardcore never-triple hitters?

[10] Written by Mark Bailey Last updated on February 7th. "What Is an Error in Baseball: Importance of the Statistic." *Baseball Bible*, 7 Feb. 2023, https://www.baseballbible.net/what-is-an-error-in-baseball/.

Returning to the slow argument, is it true that 40-year-old (and now retired) catcher Yadier Molina, who was the slowest player in the dataset, had a zero probability of hitting a triple this past season? One would have to go back six years to find a single Molina-hit triple – one of only seven over a lengthy nineteen-year career. Moreover, his sprint speed has fallen 9% since then, from a point that already would have made him the slowest player to achieve the feat in this year's class of hitters.

Ultimately, compelling cases can be made on both sides of the hardcore never-triple hitters debate. However, the randomness argument seems more believable, and therefore the spike-free models, seen as more narratively accurate, were chosen to be the focus of this analysis. Baseball is often a bizarre game. Sometimes fielding issues and careening balls buy runners so much time that they make it all the way home on a single ball hit into play. It's safer to assume that anything is possible in this context.

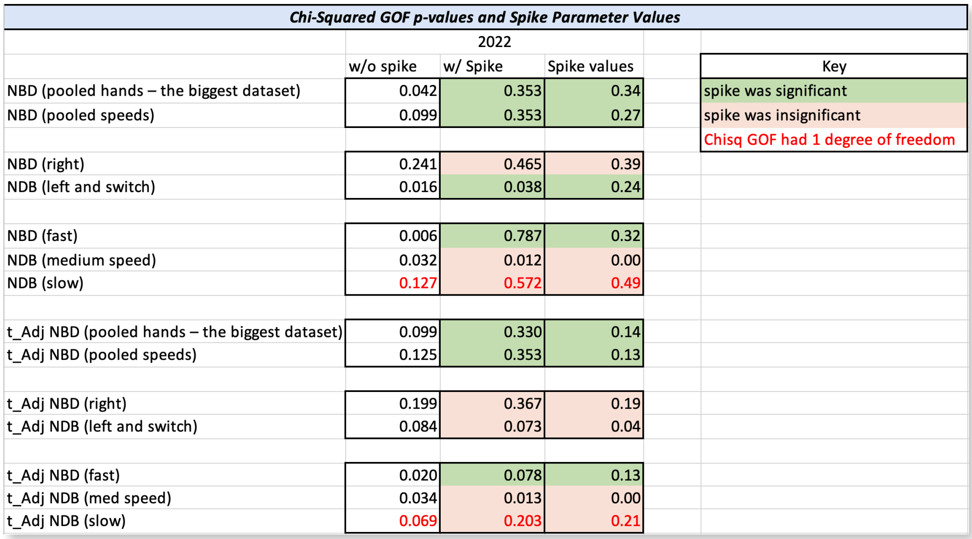
Figure 5



Perhaps more important, however, is the t-adjustment, which allows the NBD model to account for each player's number of batting attempts. While some segments performed better regarding their Chi-Squared GOF tests before introducing this data, adjusting for variations in the number of player at-bats makes more sense. Above are images of the batting-attempt distributions by segment (see *Figure 5*). All subgroups have a right skew, with the highest counts in the 0-50 range. The standard NBD acts like a t-adjusted NBD in which all players have the same number of batting attempts (one each). Clearly, this is not the case, and thus the t-adjustment is welcomed.

In terms of performance, spike-free NBD models struggled to fit the data well (see *Figure 6*), with their hardcore never-triple hitters-adjusted counterparts benefitting non-segmented models the most. One of the theoretically best-equipped subgroups to hit triples, the fast batters, was the only t-adjusted segment that significantly improved with the inclusion of a spike at zero (LRT p-value < 0.002). This segment had the smallest percentage of players without a recorded triple. Yet, even when considering their number of hitting opportunities, the model believed there were too many zeros. With these players standing out in a sea of hitters who made the task look easy, the model concluded that 13% of the subgroup (see *Figure 6*) must not be spinning their "Poisson wheels." Surely, though, if any subgroup were theoretically most capable of having all its members hit at least one triple over an infinite number of at-bats, it would be this one. Even by seemingly pure luck, all anyone with this speed would need is one well-contacted ball into a gap

Figure 6



Chi-Squared GOF p-values and Spike Parameter Values

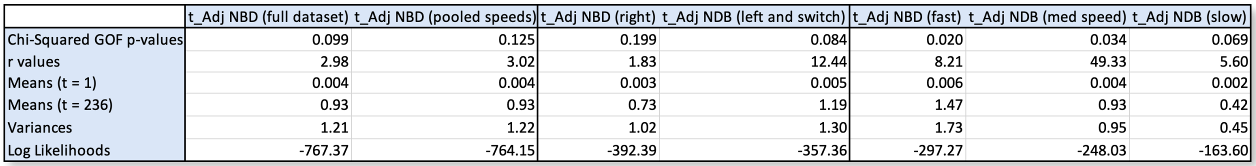
2022

	w/o spike	w/ Spike	Spike values	Key
NBD (pooled hands – the biggest dataset)	0.042	0.353	0.34	spike was significant
NBD (pooled speeds)	0.099	0.353	0.27	spike was insignificant
				Chisq GOF had 1 degree of freedom
NBD (right)	0.241	0.465	0.39	
NDB (left and switch)	0.016	0.038	0.24	
NBD (fast)	0.006	0.787	0.32	
NDB (medium speed)	0.032	0.012	0.00	
NDB (slow)	0.127	0.572	0.49	
t_Adj NBD (pooled hands – the biggest dataset)	0.099	0.330	0.14	
t_Adj NBD (pooled speeds)	0.125	0.353	0.13	
t_Adj NBD (right)	0.199	0.367	0.19	
t_Adj NDB (left and switch)	0.084	0.073	0.04	
t_Adj NBD (fast)	0.020	0.078	0.13	
t_Adj NDB (med speed)	0.034	0.013	0.00	
t_Adj NDB (slow)	0.069	0.203	0.21	

in left field. Ultimately, it's not surprising that spike models generally fit better. Still, their near total lack of significance on the segmented, and especially t-adjusted segmented models, provide greater support against their use.

From the spike-free, t-adjusted point of view, most segmented models and both variations of the larger pooled models performed better with the addition of batting-attempt information. Nevertheless, the fits were still generally poor, with only the right-hand segment NBD model performing pretty well (Chi-Squared GOF p-value = 0.199). Surprisingly, though, this right-hitting segment, along with the slow subgroup, experienced performance drop-offs after the t-adjustment. Although unintuitive, this is relatively common. Despite including highly-relevant information, sometimes more data is harder to fit, as the true story presents a more challenging landscape to model. One final point regarding the calculations of the Chi-Squared GOF tests, all rollups were performed according to the same rule: maximize the number of bins by right-truncating as far out as possible while maintaining at least 80% of expected values of 5 or more (when rounded to one decimal place).

Figure 7



	t_Adj NBD (full dataset)	t_Adj NBD (pooled speeds)	t_Adj NBD (right)	t_Adj NDB (left and switch)	t_Adj NBD (fast)	t_Adj NDB (med speed)	t_Adj NDB (slow)
Chi-Squared GOF p-values	0.099	0.125	0.199	0.084	0.020	0.034	0.069
r values	2.98	3.02	1.83	12.44	8.21	49.33	5.60
Means (t = 1)	0.004	0.004	0.003	0.005	0.006	0.004	0.002
Means (t = 236)	0.93	0.93	0.73	1.19	1.47	0.93	0.42
Variances	1.21	1.22	1.02	1.30	1.73	0.95	0.45
Log Likelihoods	-767.37	-764.15	-392.39	-357.36	-297.27	-248.03	-163.60

With a reasonably small dataset and even smaller subgroups, it's not entirely surprising that the models fit the data poorly (see *Figures 7 & 8*). Nevertheless, both segmentations were highly significant (see *Figure 9*). As expected, batting hands and sprint speeds are very relevant covariates for modeling triples. Belief in these underlying stories inspires enough confidence to analyze further the segmented player propensity distributions (their lambda distributions). While all subgroups were technically homogeneous (with interior modes and r values greater than 1), their varying levels of homogeneity highlight interesting segment differences (see *Figure 10*).

Figure 8

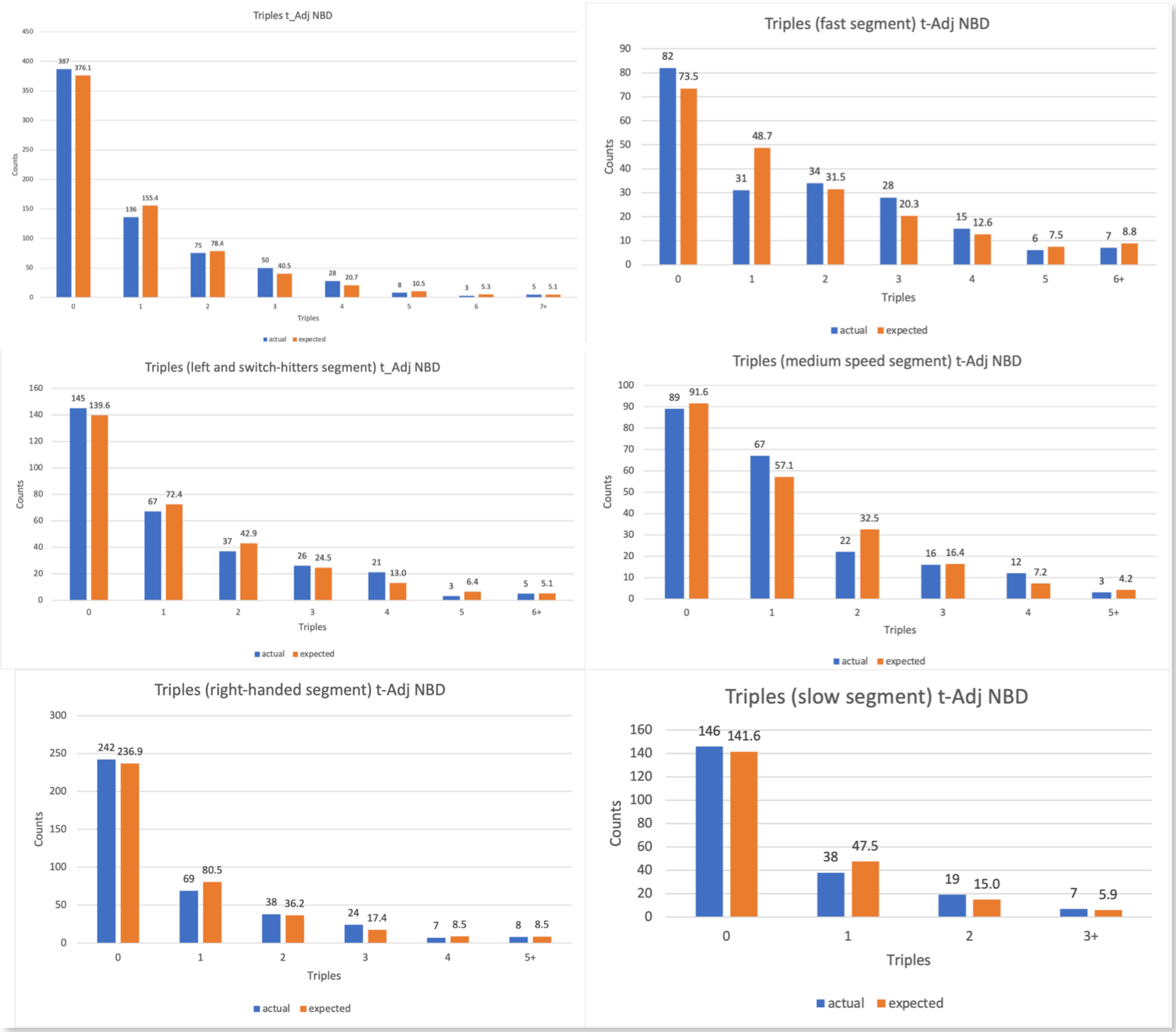
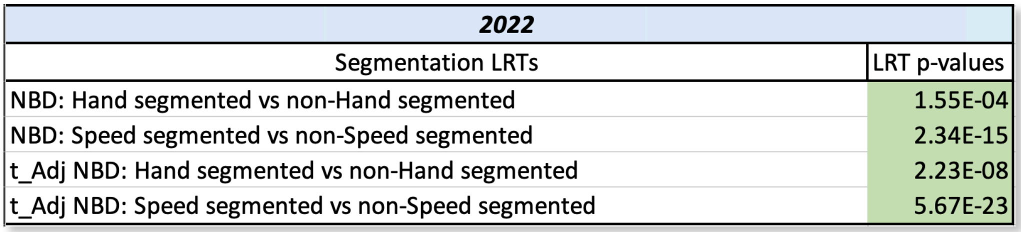


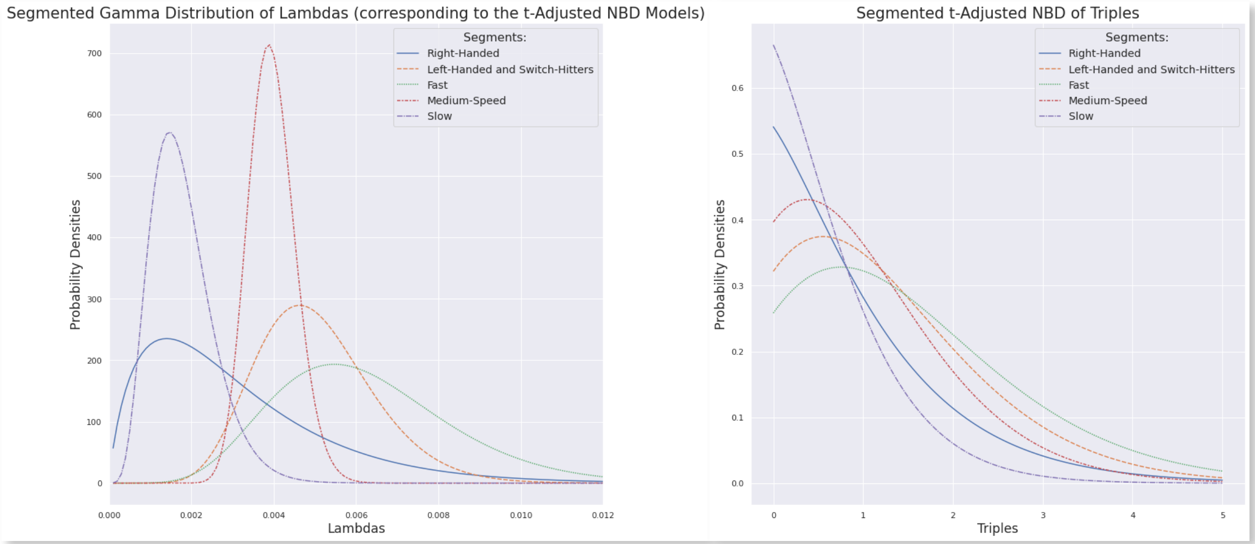
Figure 9



2022	
Segmentation LRTs	LRT p-values
NBD: Hand segmented vs non-Hand segmented	1.55E-04
NBD: Speed segmented vs non-Speed segmented	2.34E-15
t_Adj NBD: Hand segmented vs non-Hand segmented	2.23E-08
t_Adj NBD: Speed segmented vs non-Speed segmented	5.67E-23

Right-handed hitters were the least homogeneous of the subgroups, containing a high concentration of triple-hitting propensities just beyond zero with a relatively long right tail. Intuitively, this means that even though a randomly selected right-handed batter is likelier to hit triples at a low rate, a decent number of players in this category nevertheless excel at reaching third. Such batters might be exceedingly fast and uncommonly skilled at pushing the ball to right field or using their speed to leg out shots to center.

Figure 10



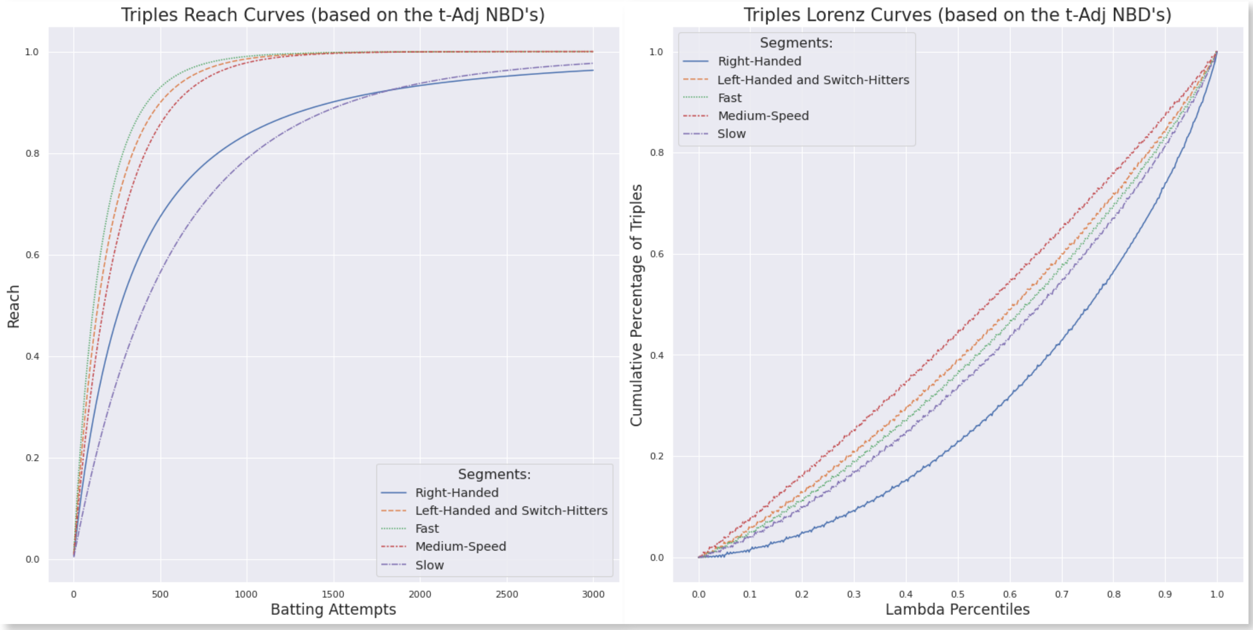
Slow players were the next least homogeneous, with a much higher clustering of propensities around its segment-low mean of 0.002 (when t=1). Despite a large frequency at this point, the subgroup still maintained more heterogeneity than many others due to a disproportionately thick (in relation to the rest of its distribution) right tail. It’s much more difficult for slow players to hit triples than any other subgroup, so when some players do it with even a moderate propensity (likely in part because they are excellent at pulling the ball to right field), this translates into a greater level of heterogeneity.

On the opposite side of the speed spectrum, the fast subgroup registered as the next least homogeneous. While slow players reduced their r metric of homogeneity by virtue of so many struggling to hit for any triples (69.5% recorded 0), the quickest players in baseball didn't rank as the most homogeneous, likely due to variations in batting success being amplified by their elite speed. Theoretically, these players are all fast enough to record lots of triples. However, all that speed is useless in converting at-bats into three-baggers if these players aren't very good at hitting the ball into right or, at the very least, deep center field. These high speeds are essentially acting as a large triples-multiplier of one's ability to hit baseballs in this specific manner. Players with both ingredients have the highest propensities to hit triples, while those lacking these batting skills might have lambdas more closely resembling people from the medium-speed subgroup.

In many ways, the left-handed/switch-hitter segment's homogeneity story mirrors the fast subgroup's. While speedsters' propensities were likely dispersed primarily according to their more variable ability to hit the ball hard into right and center field, left and switch-hitters differ in their ability to quickly round the bags. However, these contrasts probably aren't magnified to the same extent. There would conceivably be more variation within the left/switch-hitter subgroup's abilities to pull the ball consistently well to right field than in the speeds of the elite third of baseball sprinters. Recalling that switch-hitters also bat from the right at times further supports this argument. In other words, the right field-hitting "triples-multiplier" theoretically shouldn't be as large as the fast subgroup's more uniform speed equivalent. If this hand-segmentation had been replaced by a metric of players' true abilities to hit balls hard into right field (such as established previous success rates), this would be a different story. However, player-handedness serves only as a proxy of this tendency (there's more noise). Thus, it's unsurprising that the left/switch-hitter subgroup is more homogeneous.

Finally, the medium-speed players reign supreme in this contest of homogeneity. These players are the definition of average when it comes to hitting triples (in fact, their mean rate actually equals the population average). They lack the top-tier speed that serves as a multiplier of one's hitting abilities in the fast subgroup. Similarly, they aren't slow enough as a whole that some players' abilities to hit for even a minute number of triples would produce a particularly stark contrast with everyone else's would-be near-zero propensities. Consequently, these highly similar medium-speed players have the most extensive spike-shaped Gamma distribution of triple-hitting propensities (lambdas).

Figure 11



Inter-segment homogeneity differences also manifest within the reach and Lorenz curves (see *Figure 11*). After a few at-bats (10), the fast, left/switch, and medium-speed segments jump up their reach curves the highest in this order, leaving behind the righties and, in last, the slow players. Yet, as hitters accrue more BAs, the slow segment's reach is eventually predicted to surpass the more heterogeneous right-handed batters. Likewise, the medium-speed subgroup ultimately overtakes both the fast and left/switch players. The point is that given sufficient time, more homogeneous groups will beat out the more heterogeneous groups, which flatten off earlier.

While homogeneity ultimately wins out in the reach curve race, heterogeneity shines in the Lorenz curve (see *Figure 8*), as a larger percentage of productivity comes from a smaller portion of the population when holding the number of batting attempts equal (a theoretical analysis). Since the medium-speed players are very similar, the relationship between the anticipated cumulative percentage of triples hit and the lambda percentiles of individuals' propensities to hit them is nearly linear. The bottom 20% of medium-speed hitters are expected to produce 16.2% of all triples in the subgroup, whereas the top 20% are predicted to account for a not-too-dissimilar 24.1%. On the other end of this spectrum, the least homogeneous right-handed hitters have a more bowed Lorenz curve. The bottom 20% are anticipated to generate just 4.7% of the segment's triples, whereas the top 20% would theoretically make up a much larger 43.5%.

In conclusion, although the preferred segmented t-adjusted NBD models failed to fit the data well, their stories were compelling enough to warrant further investigation into the different batting-hand and speed-segmented underlying triple-hitting propensities. In the future, a non-stationarity analysis over multiple seasons, with cross-segmentations on the currently siloed player speed and batting handedness division points, would be fascinating. Considering the limitations of this model's smaller sample size (692 players segmented by hand and 622 by speed), perhaps additional seasons would allow for more granular segmentations and produce better-performing models.