A Consideration of Probability and Runs Scored Major League Baseball Extra-Inning Games

Paul A. Hodgetts

02/03/2021

Abstract

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Nisl suscipit adipiscing bibendum est ultricies integer quis auctor elit. Facilisis leo vel fringilla est. Dignissim suspendisse in est ante in. Condimentum vitae sapien pellentesque habitant. Dui vivamus arcu felis bibendum ut tristique et egestas. Orci sagittis eu volutpat odio facilisis mauris. Vulputate sapien nec sagittis aliquam malesuada bibendum arcu. Ut sem viverra aliquet eget. Ullamcorper eget nulla facilisi etiam dignissim diam quis enim. Rhoncus est pellentesque elit ullamcorper dignissim. Tellus mauris a diam maecenas. Et magnis dis parturient montes nascetur ridiculus mus. Eu sem integer vitae justo. Semper feugiat nibh sed pulvinar proin. Quisque non tellus orci ac.

Introduction

The outbreak of COVID-19 and the subsequent pandemic led to a plethora of questions and concerns in the sporting world, including whether leagues would commit to a 2020 season. For those leagues that did decide to host a 2020 seasons, various protocols were required to ensure the health and safety of the athletes and staff. For instance, the National Hockey League (NHL) Implemented bubbles with all teams within the Western Conference playing within Edmonton, Alberta and all teams within the Eastern Conference within Toronto, Ontario (Gatto, 2020). In a similar move, the National Basketball Association (NBA) established a bubble in Orlando, Florida within which teams could play out the season (Haislop, 2020). However, unlike the use of a bubbled league like the NHL and NBA, Major League Baseball (MLB) permitted teams to play games within their own stadiums, excluding the Toronto Blue Jays who were denied access to play within Canada by the Canadian federal government (McNamara, 2020; Wagner, 2020). In choosing this approach, MLB implemented other policies such as no spitting, masks being required in the dugout and bullpen, and no saunas, and twice-a-day temperature and symptom checks to name a few (Wagner, 2020). One such policy was to also introduce a new rule regarding extra-inning games, tied games that go beyond the regulation nine innings, in hopes of shortening the exposure experienced by players between teams (Allen, 2020). The rule was that if at the completion of the regulation innings a game was tied, each team would begin the subsequent half-inning with the last player to make an out on second base (Allen, 2020). As a rule change to a sport or game should ensure the fairness of the playing field, this paper looks to examine this rule change regarding whether it provides an advantage to the away team in extra-inning games through the probabilities of runs scored based on the state of events in a half-inning. Moreover, it considers whether extra-inning games in general provide an advantage to the away team through the probabilities of runs scored based on the state of events in a half-inning, and discusses strategy within the context of extra-inning games.

Data

Location

This analysis uses game-log and play-by-play data from the 2000 MLB season to the 2020 MLB season. Game-logs and play-by-play files were obtained free of charge from are copyrighted by Retrosheet. Interested parties may contact Retrosheet at "www.retrosheet.org".. The play-by-play files were accessed using the `parse_retrosheet_pbp()` function from GitHub user "beanumber", with the process described by Marchi et al. in 'Appendix A' of 'Analyzing Baseball Data with R' (2019c).

Other data files include the fields dataset, which provides the Retrosheet event headers. This can be accessed from from "the baseball_R GitHub repository maintained by user maxtoki".

Missing Values

Regarding game-log data, four missing values were found in the regulation length games and two missing values were found in extra-inning games due to tie-games creating neither a winner nor loser. These values were removed from their respective datasets leaving 45,283 observations in regulation length games and 4,197 in extra-inning games.

Play-by-play files were also examined for missing values; however, all missing values belonged to event-type variables (e.g. the play on runner on second), so these values were not removed as doing so would create issues within the data regarding analysis.

Exploratory Analysis

An exploratory analysis comparing home wins against visitor wins for both extra-inning games and regulation length games for all seasons revealed that over all seasons the home team won more games, for both extra-inning games and regulation games (see Figure 1). Breaking down regulation length games by season also showed that across all seasons the home team won more games than the visiting team (see Figure 2 and Table 1). Additionally, breaking down each extra-inning game by season, generally shows the same pattern of the home team winning more games than the visitor, with the home team winning more games in 15 of the 21 seasons (see Figure 2 and Table 2). However, in the 2000 and 2012 seasons both the home team and visitor won an equal number of extra-inning games at 101 and 96 games apiece respectively (see Figure 2 and Table 2). Meanwhile, the visitor won more extra-inning games than the home team in 2001, 2014, 2019, and 2020 (see Figure 2 and Table 2).

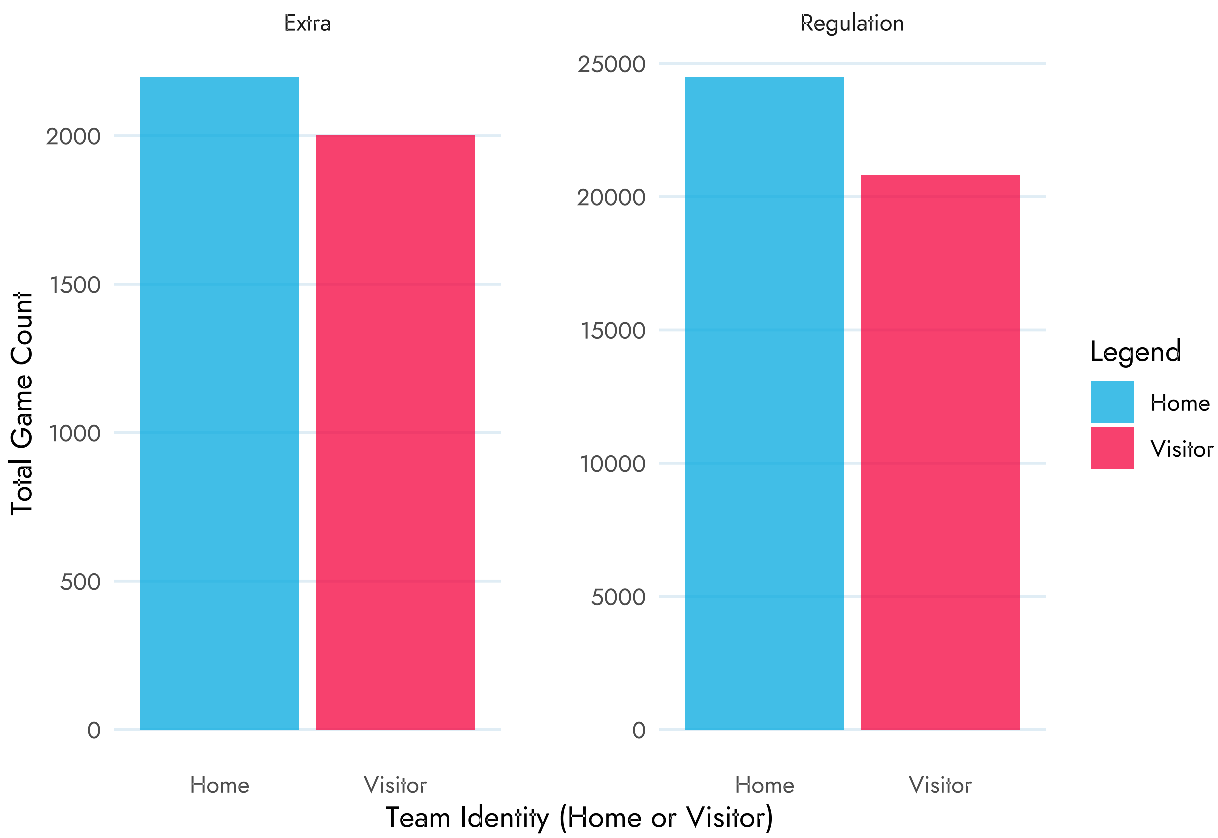

Figure 1: Home vs. Visitor Wins for Extra-Inning Games and Regulation Length Games for MLB Seasons 2000-2020

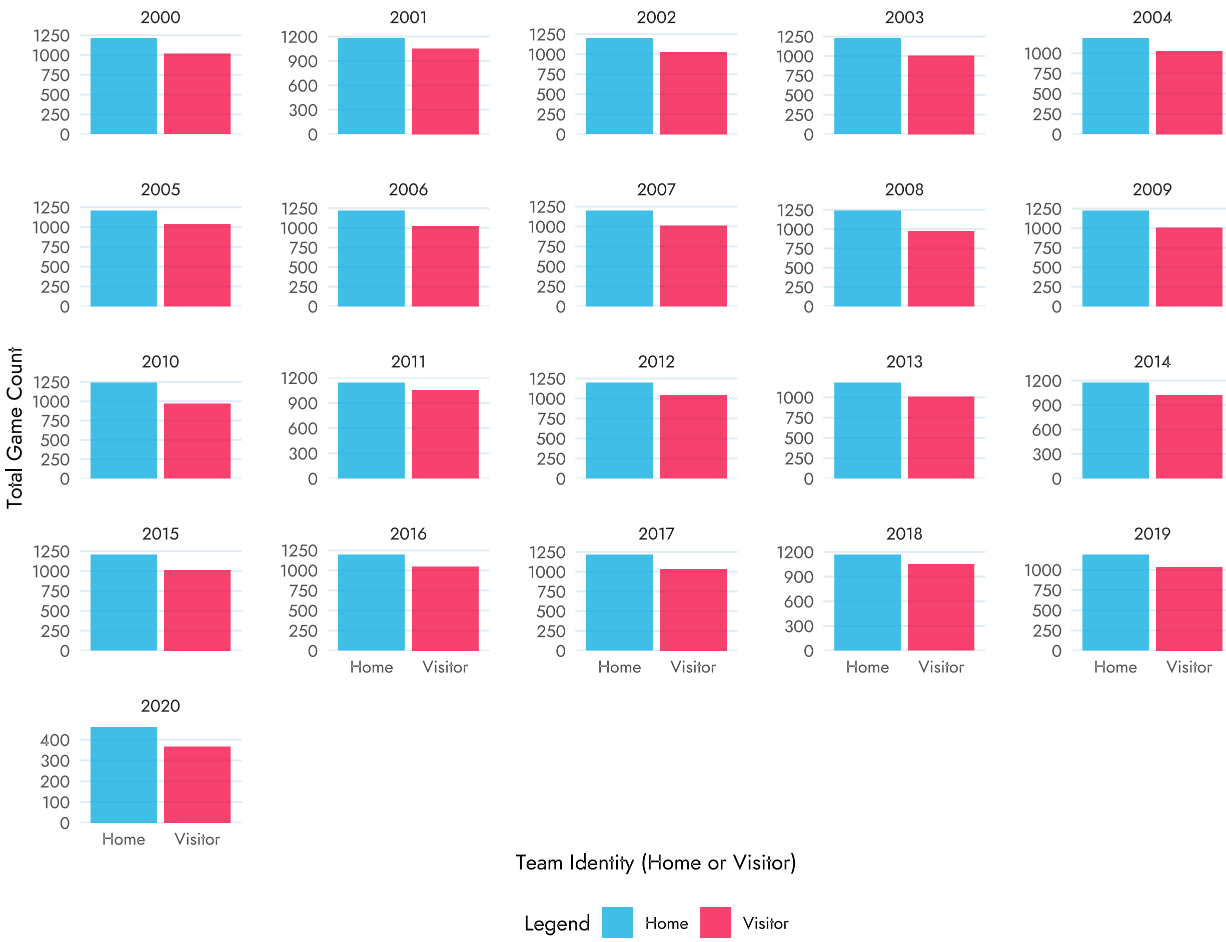

Figure 2: Home vs. Visitor Wins for Regulation Length Games for MLB Seasons 2000-2020

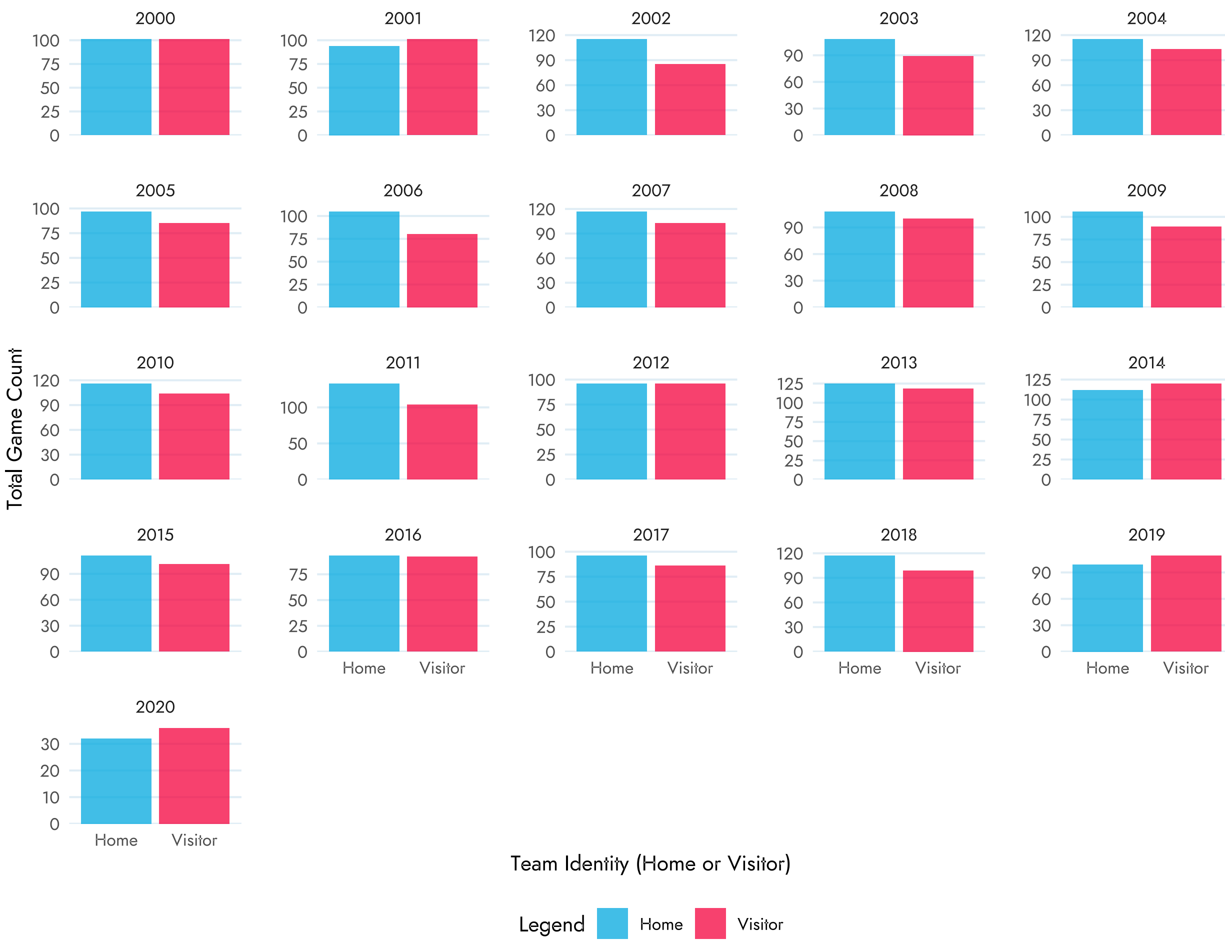

Figure 3: Home vs. Visitor Wins for Extra-Inning Games for MLB Seasons 2000-2020

Table 1: Regular Inning Win Count for Home and Visitor by Season

Home		Visitor	
Wins	Season	Wins	Season
1211	2000	1015	2000
1179	2001	1054	2001
1199	2002	1026	2002
1227	2003	1005	2003
1184	2004	1026	2004
1209	2005	1039	2005
1222	2006	1022	2006
1201	2007	1010	2007
1243	2008	977	2008
1227	2009	1008	2009
1242	2010	968	2010
1143	2011	1049	2011
1199	2012	1039	2012
1182	2013	1006	2013
1176	2014	1022	2014
1205	2015	1012	2015
1194	2016	1048	2016
1215	2017	1033	2017
1166	2018	1049	2018
1187	2019	1034	2019
462	2020	368	2020

Table 2: Extra Inning Win Count for Home and Visitor by Season

Home		Visitor	
Wins	Season	Wins	Season
101	2000	101	2000
94	2001	101	2001
115	2002	85	2002
108	2003	89	2003
115	2004	103	2004
97	2005	85	2005
105	2006	80	2006
117	2007	103	2007
108	2008	100	2008
106	2009	89	2009
116	2010	104	2010
133	2011	104	2011
96	2012	96	2012
125	2013	118	2013
112	2014	120	2014
111	2015	101	2015
93	2016	92	2016
96	2017	86	2017
117	2018	99	2018
99	2019	109	2019
32	2020	36	2020

Analysis

Runs-Expectancy Matrix

To begin, a run expectancy matrix was generated to calculate the average number of runs scored from the different base-out states. In baseball, there are three possible states of outs before an inning is over, e.g., zero, one, or two outs, and each base (i.e., first, second, and third) can either be in a state of being occupied or not occupied. Thereby giving 24 possible base-out states (8 base states x 3 out states = 24 base-out states). This was calculated for each MLB season from 2000 to 2020, as well as for the seasons combined. Table 3 shows the runs-expectancy matrix for the total combined seasons. To read this table, the the first column represents a base state with the next three columns showing an out state. Base states can be read as follows, 0 represents an unoccupied base with 1 representing an occupied base. First base is represented by the first numeral on the left, with second base being the middle numeral, and third base being the right-most numeral. For example, reading the table at the row beginning with 010 means a runner on second with the following average runs in the three out states. Continuing along this row, it can be seen that in this state with no outs an average number of 1.14 runs are scored, which drops to 0.69 average runs with one out, and subsequently 0.33 runs with two outs.

Table 3: Runs Expectancy Matrix for All Seasons (2000-2020) Combined

	0 outs	1 out	2 outs
000	0.51	0.27	0.10
001	1.42	0.97	0.37
010	1.14	0.69	0.33
011	2.01	1.41	0.58
100	0.89	0.53	0.23
101	1.80	1.18	0.50
110	1.49	0.92	0.44
111	2.32	1.58	0.77

In addition to the average runs scored, the variance and standard deviation of runs scored from a base-out state were also calculated and are displayed in Table 4. Here it can be seen that the base-out state with the greatest standard deviation is the bases loaded and no one out (111 0), with a variance of $\sigma^2 = 3.34$ and a standard deviation of $\sigma = 1.83$. Additionally, the base-out state with the lowest standard deviation is no one on and two outs (000 2), with a variance of $\sigma^2 = 0.19$ and standard deviation of $\sigma = 0.44$. Notably, the base-out state of runner on second and no one out is towards the middle of the base-out states in terms of standard deviation at position 11 of 24 and with a variance of $\sigma^2 = 1.75$ and standard deviation of $\sigma = 1.32$ (see Table 4).

Table 4: Distribution and Central Tendency for Expected Runs for All Seasons (2000-2020) Combined

State	Mean	Median	Variance	SD	Range
000 2	0.1046424	0	0.1932411	0.4395920	0-9
100 2	0.2302677	0	0.4787442	0.6919134	0-9
010 2	0.3271166	0	0.5500225	0.7416350	0-10
000 1	0.2710119	0	0.5524525	0.7432715	0-12
001 2	0.3675043	0	0.5692348	0.7544765	0-11
110 2	0.4414346	0	0.9641316	0.9819020	0-11
101 2	0.5005371	0	1.0014179	1.0007087	0-12
000 0	0.5073049	0	1.0712216	1.0349984	0-14
001 1	0.9651565	1	1.0829098	1.0406295	0-11
100 1	0.5301370	0	1.1275103	1.0618429	0-12
010 1	0.6884575	0	1.1835046	1.0878900	0-12
011 2	0.5810501	0	1.2328890	1.1103553	0-10
001 0	1.4167235	1	1.5863636	1.2595093	0-11
010 0	1.1377177	1	1.7483712	1.3222599	0-14
101 1	1.1820328	1	1.7944072	1.3395549	0-12
111 2	0.7683975	0	1.8166367	1.3478267	0-11
100 0	0.8898628	0	1.8304100	1.3529265	0-14
110 1	0.9208802	0	1.9351109	1.3910827	0-12
011 1	1.4071204	1	2.0222704	1.4220655	0-12
101 0	1.8008865	1	2.3904744	1.5461159	0-13
011 0	2.0092898	2	2.4080953	1.5518039	0-13
110 0	1.4887929	1	2.7390837	1.6550177	0-14
111 1	1.5750796	1	2.7714777	1.6647756	0-11
111 0	2.3238272	2	3.3451838	1.8289844	0-13

However, while these tables show the average expected runs from a given base-out state, there is also the question as to the probability of a run actually scoring given a base-out state. Table 5 shows the mean of value of at least one run scored given each base-out state. Looking at this table, the base-out state with the greatest probability of one or more runs scoring is the bases loaded and no outs (111 0) with a probability of 0.86, and the lowest probability of one or more runs scoring is no one on and two outs (000 2) with a probability of 0.07. Looking to the base-out state of a runner on second and no outs (010 0), or the starting state for each extra-inning game in the 2020 MLB season, it can be seen that at least one run scoring given that state has a probability of 0.62. This is compared to the standard state of beginning of an inning with no runners and no outs (000 0), which has a probability of 0.28 for at least one or more runs scoring from that state.

Table 5: Probability of Scoring at least One Run from a Given Base-Out State

State	Run Probability
111 0	0.8599573
101 0	0.8545246
011 0	0.8530866
001 0	0.8394893
011 1	0.6786357
111 1	0.6611908
001 1	0.6556419
101 1	0.6377960
110 0	0.6198359
010 0	0.6184847
100 0	0.4203736
110 1	0.4132753
010 1	0.4012692
111 2	0.3194076
000 0	0.2766934
101 2	0.2739754
100 1	0.2694704
011 2	0.2596885
001 2	0.2578975
110 2	0.2251555
010 2	0.2174605
000 1	0.1625228
100 2	0.1286415
000 2	0.0708171

State Transitions

A half-inning of baseball lasts until a third out is made and up to that point the game moves between the various possible base-out states. For example, a regulation half-inning could progress from the first batter as follows: no runners and no outs (000 0), to a runner on first with no outs (100 0), to no runners and two outs (000 2), to runner on second and two outs (010 2), to finally three outs and the end of the half-inning. Using the current base-out and the subsequent new base-out state, a transition table was produced to calculate the frequency at which one state followed another. From this transition table a proportional table was then produced to view the proportion at which one state moved to another. This table is displayed in Table 5. To read this table, each row and column represents a state, with the values in each cell representing the proportion of moving from the base-out state of that row to the corresponding column base-out state.

Table 6: Proportional Matrix for Transitions Between Base-Out States

	000 0	000 1	000 2	001 0	001 1	001 2	010 0	010 1	010 2	
000 0	0.0306306	0.6731739	0.0000000	0.0056567	0.0000000	0.0000000	0.0512308	0.0000000	0.0000000	0
000 1	0.0000000	0.0275954	0.6808377	0.0000000	0.0051448	0.0000000	0.0000000	0.0477728	0.0000000	0
000 2	0.0000000	0.0000000	0.0277363	0.0000000	0.0000000	0.0045842	0.0000000	0.0000000	0.0470775	0
001 0	0.0244576	0.2237673	0.0039448	0.0062130	0.4013807	0.0000000	0.0505917	0.0009862	0.0000000	0
001 1	0.0000000	0.0231575	0.2100880	0.0000000	0.0071013	0.3479623	0.0000000	0.0500483	0.0055870	0
001 2	0.0000000	0.0000000	0.0223231	0.0000000	0.0000000	0.0048948	0.0000000	0.0000000	0.0432912	0
010 0	0.0225424	0.0026140	0.0062707	0.0051840	0.2698182	0.0000000	0.0480659	0.3767733	0.0000000	0
010 1	0.0000000	0.0243515	0.0036774	0.0000000	0.0058390	0.1865694	0.0000000	0.0514207	0.4354393	0
010 2	0.0000000	0.0000000	0.0223416	0.0000000	0.0000000	0.0061529	0.0000000	0.0000000	0.0566706	0
011 0	0.0228550	0.0028094	0.0023538	0.0056948	0.1519362	0.0037965	0.0526196	0.0897494	0.0024298	0
011 1	0.0000000	0.0182787	0.0037487	0.0000000	0.0060413	0.1292831	0.0000000	0.0482992	0.0863746	0
011 2	0.0000000	0.0000000	0.0202004	0.0000000	0.0000000	0.0054731	0.0000000	0.0000000	0.0507654	0
100 0	0.0291043	0.0004954	0.1210608	0.0054449	0.0025743	0.0000000	0.0144307	0.1127356	0.0000000	0
100 1	0.0000000	0.0293340	0.0006474	0.0000000	0.0061299	0.0027516	0.0000000	0.0158498	0.0833454	0
100 2	0.0000000	0.0000000	0.0290452	0.0000000	0.0000000	0.0068369	0.0000000	0.0000000	0.0226232	0
101 0	0.0294789	0.0009030	0.0845594	0.0061082	0.0025495	0.0105168	0.0143942	0.0487598	0.0033994	0
101 1	0.0000000	0.0270270	0.0006592	0.0000000	0.0065431	0.0031251	0.0000000	0.0156742	0.0633560	0
101 2	0.0000000	0.0000000	0.0246450	0.0000000	0.0000000	0.0073991	0.0000000	0.0000000	0.0224589	0
110 0	0.0284717	0.0004768	0.0001526	0.0055494	0.0012205	0.0912506	0.0134063	0.0025363	0.0109653	0
110 1	0.0000000	0.0283132	0.0006447	0.0000000	0.0062321	0.0016118	0.0000000	0.0161821	0.0039757	0
110 2	0.0000000	0.0000000	0.0253608	0.0000000	0.0000000	0.0078014	0.0000000	0.0000000	0.0238826	0
111 0	0.0277525	0.0005592	0.0002097	0.0057323	0.0021671	0.0750786	0.0144006	0.0033555	0.0032856	0
111 1	0.0000000	0.0274306	0.0008009	0.0000000	0.0071508	0.0016018	0.0000000	0.0147593	0.0033180	0
111 2	0.0000000	0.0000000	0.0256935	0.0000000	0.0000000	0.0093706	0.0000000	0.0000000	0.0248564	0
	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0

Runs scored $RUNS$ is equal to difference between the sum of runners $N_{runners}$ and outs O before (b) the event plus one and the number of runners $N_{runners}$ plus outs O after (a) after the event.

$$RUNS = (N_{runners}^{(b)} + O^{(b)} + 1) - (N_{runners}^{(a)} + o^{(a)})$$

References

Gatto, T. (2020, August 14). *NHL bubble, explained: A guide to the hub city rules, teams & schedule for Edmonton, Toronto*. Sporting News. https://www.sportingnews.com/us/nhl/news/nhl-bubble-hub-city-rules-teams-schedule-edmonton-toronto/72k8vc0u630k19xalra66xa3c

Haislop, T. (2020, August 26). *NBA bubble explained: A complete guide to the rules, teams, schedule & more for Orlando games*. Sporting News. https://www.sportingnews.com/us/nba/news/nba-bubble-rules-teams-schedule-orlando/zhap66a9hcwq1khmcex3ggabo

Marchi, M., Albert, J., & Baumer, B. S. (2019a). Chapter 5: Value of plays using run expectancy. In, *Analyzing baseball data with R* (2nd ed.), (pp. 111-135). CRC Press

Marchi, M., Albert, J., & Baumer, B. S. (2019b). Chapter 9: Simulation. In, *Analyzing baseball data with R* (2nd ed.), (pp. 201-226). CRC Press

Marchi, M., Albert, J., & Baumer, B. S. (2019c). Appendix A: Retrosheet files reference. In, *Analyzing baseball data with R* (2nd ed.), (pp. 293-301). CRC Press

Table 7: Probability of Next Base-Out State After One At-Bat for Man on Second no Outs

state	new_state	prob
010 0	010 1	0.3767733
010 0	001 1	0.2698182
010 0	110 0	0.1053250
010 0	101 0	0.0980850
010 0	010 0	0.0480659
010 0	100 0	0.0470820
010 0	000 0	0.0225424
010 0	100 1	0.0138338
010 0	000 2	0.0062707
010 0	001 0	0.0051840
010 0	011 0	0.0044057
010 0	000 1	0.0026140
010 0	001 2	0.0000000
010 0	010 2	0.0000000
010 0	011 1	0.0000000
010 0	011 2	0.0000000
010 0	100 2	0.0000000
010 0	101 1	0.0000000
010 0	101 2	0.0000000
010 0	110 1	0.0000000
010 0	110 2	0.0000000
010 0	111 0	0.0000000
010 0	111 1	0.0000000
010 0	111 2	0.0000000
010 0	3	0.0000000

McNamara, A. (2020, July 24). *Toronto Blue Jays to play majority of 2020 home games in Buffalo.* CBS News. https://www.cbsnews.com/news/toronto-blue-jays-home-games-buffalo-2020-season/

Wagner, J. (2020, June 24). *Baseball's New Rules: No Spitting, No Arguing, and Lots of Testing.* The New York Times. https://www.nytimes.com/2020/06/24/sports/baseball/mlb-coronavirus-rules.html

Table 8: Probability of Next Base-Out State After Two At-Bats for Man on Second no Outs

state	new_state	prob
010 0	001 2	0.1749070
010 0	010 2	0.1685166
010 0	100 1	0.1075496
010 0	101 1	0.1017980
010 0	110 1	0.0956944
010 0	000 2	0.0743758
010 0	010 1	0.0620901
010 0	110 0	0.0334917
010 0	000 1	0.0325363
010 0	111 0	0.0260449
010 0	011 1	0.0211149
010 0	001 1	0.0204333
010 0	100 2	0.0190616
010 0	101 0	0.0138316
010 0	3	0.0118905
010 0	011 0	0.0092679
010 0	000 0	0.0092620
010 0	100 0	0.0087972
010 0	010 0	0.0074626
010 0	001 0	0.0018740
010 0	011 2	0.0000000
010 0	101 2	0.0000000
010 0	110 2	0.0000000
010 0	111 1	0.0000000
010 0	111 2	0.0000000

Table 9: Probability of Next Base-Out State After Three At-Bats for Man on Second no Outs

state	new_state	prob
010 0	3	0.3339465
010 0	100 2	0.1220218
010 0	110 2	0.0697820
010 0	101 2	0.0684754
010 0	010 2	0.0663450
010 0	110 1	0.0641391
010 0	000 2	0.0395578
010 0	111 1	0.0367266
010 0	101 1	0.0301678
010 0	001 2	0.0297487
010 0	100 1	0.0225766
010 0	011 1	0.0202370
010 0	000 1	0.0186317
010 0	010 1	0.0171222
010 0	011 2	0.0168674
010 0	111 0	0.0124475
010 0	110 0	0.0072185
010 0	001 1	0.0070534
010 0	101 0	0.0041257
010 0	100 0	0.0033659
010 0	011 0	0.0032048
010 0	000 0	0.0030497
010 0	010 0	0.0025658
010 0	001 0	0.0006230
010 0	111 2	0.0000000