

Statistics on the 2021 Philadelphia Phillies

A problability and statistics report

Joshua Distefano & Joe DeLizza | CSCI 3327 | 4/28/2022

# Background

The Philadelphia Phillies is a major league baseball team that is in the National League east division. The team was building in the year 2021 after the lackluster year before in 2020. These statistics that are used for this report are in the 2021 season. Since then, some players have been traded and released so the stats sheet for this year might look different than the one compared in this report. Some things to clarify are the acronyms that are present on the statistics sheet. The term RBI’s (runs batted in) means that the batter put the ball in play and scored a run because of it. OBP (on base percentage) is the percentage the batter reached a base in their plate appearances. The AVG that is displayed is the rate at which the batter gets a hit compared to his plate appearances. While the AVG shown on the pitcher’s side stands for the rate in which an opposing batter gets a hit against that pitcher in question. Finally, ERA (earned run average) means the average number of runs allowed within nine innings of play.

# Calculations

When comparing the batting averages of all of the fielders for the Philadelphia Phillies, the averages can seem to meld together. All of the averages are close to one another. However, by using a relative frequency histogram we can see how many fielders have a batting average similar to one another. As shown on the left, the averages of the fielders for the Phillies hover around the .240-.260 range. Two players are on the lower end of the spectrum of batting averages with a .200-.220 batting average. Another two players are better than the team average with .260-.280 batting average. Rounding it off, only two players are better than the .260-.280 range and one is in the middle of a .220-.240 batting average. Overall, this mean that most of the team about a quarter of the time will be able to get a hit and get on base for their team.

As for what type of hits the players achieve, this can be manipulated by using numeric methods of measurements to show some consistencies. The average or mean of doubles, triples, and home runs can be found by using the following formula: . By adding up all the values and the set for each category and then dividing by the number of values for each category. When doing so, the mean of doubles hit is 22.4, 1.7 triples hit, and eighteen home runs hit. To find the middle value of the categories or the median, each category must be organized from smallest to largest. Once that occurs, the value can be picked out by going through the data sets and picking the value in the middle for sets with an odd number of values. The median for doubles hit, triples hit, and home runs hit is 24.5, 1.5, and 15.5, respectively. To find the variance of these sets of data (or the distance that the data sets stray from the mean) this formula will be used: . This means that all of the data values will be subtracted by the mean of each set of data then squared. Afterwards, that value will be multiplied by one divided by the total number of values minus one. After the calculations, the variance for doubles, triples, and home runs hit are 100.93, 1.78, and 84.88, respectively. For the final bit of data manipulation, the standard deviation of the doubles, triples, and home runs hit can be found by taking the square root the variance from each data set found earlier. In doing so, the standard deviation for doubles hit is 10.04, triples hit is 1.33, and home runs hit is 9.21.

A baseball player swinging a bat

Description automatically generated with medium confidenceWhen it comes to the players being up to the plate, they must concentrate and learn to look for patterns. For example, if a pitcher is known to throw what is known as “down and away” from the home plate, then the batter should know to not swing and get a ball instead of a strike. With this in mind let us ask a question then. In the span of three pitches, what is the probability that the count could be one ball and two strikes? Well first we can set up simple events to help visualize all events. These events are: . We can see that for each chance occurring naturally it would be a one-eighth chance of happening since there are eight total possibilities. The only possible ways to get one ball and two strikes are in P3, P5, and P7. So that is the three separate chances all added up together (1/8 + 1/8 + 1/8) which would make a 3/8 chance for the count to be one ball and two strikes within three pitches. If a player gets the most hits for their team, they can receive a bonus in their contract. Other players can still receive bonuses albeit less than the lead guy for being in 2nd and 3rd position for hits made in the season.

Let us say that all ten people in the fielder sheet are eligible for receive this bonus. What is the number of ways that each player can be 1st, 2nd, and 3rd respectively? A permutation can be used here to figure out the number of ways that this could happen. Permutations are the ordered arrangement of objects and using them can determine the number of ways that the objects can be ordered in a particular order. The formula for a permutation is which means the factorial of the sample size divided by the factorial of the number of objects in a set minus the number of positions being looked for. In this case, the number of players in the fielder set is ten so that will be the *n* value. The number of positions being looked for are three with those being 1st, 2nd, and 3rd position. The problem will now be set up like so: . When we solve this problem, it will come to (10 \* nine \*8) which comes to 720 combinations of players being chosen 1st, 2nd, and 3rd place for their bonus. If we want to figure out if someone can get a hit out of the whole team, we can use combinations.

A combination is the number of unordered subsets based on the number of objects being chosen out of the main set of values without replacement. The formula for a combination is set up like this here: . In this data set we can ask the question: out of all players on the team, what is the number of ways three players can get an RBI in a game of baseball? So, there are ten players in this data which include the starting nine and a possible designated hitter. The problem here is asking for the number of combinations of three people scoring an RBI. The combination formula will be set up like so: . This will turn into factorial of ten divided by three factorial times seven factorial. Doing the final calculations reveal that there are 120 separate ways for the nine players and designated hitter to achieve three RBIs in one game of baseball.

A baseball player throwing a ball

Description automatically generatedWhen choosing the starting pitcher for a game, it can be like a chess match between the two teams. The teams will compare the batting styles and batting sides that the other team has in order to pick the correct pitcher for the game. In some cases, there can be two starters to help rest the main starting pitcher for other games. For this case let us say that in order to remove bias from choosing the starting pitcher, the chosen pitcher is at random in the pool of starters available. These pitchers are Kyle Gibson, Aaron Nola, Zach Eflin, Zach Wheeler, and Ranger Suarez. Out of these pitchers, Ranger Suarez and Zack Wheeler bat left. The rest bat on the right side. What would be the probability distribution for selecting two pitchers from right-handed batting pitchers to left-handed batting pitchers? For this scenario, the probability distribution of a random variable can be used. This means that the sum of all probabilities in a set of sample points will be able to equal one. A formula for this looks like the following: for all *y*. Back to the problem at hand, let us take the combination of the five starting pitchers to the two we want to select . This combination equals ten so that means the set of sample points will be ten values. All combinations of players selected have an equally likely 1/10 chance of occurring. For choosing none of the left-handed batters, the equation would look like so: . Solving this equation leads to a three-fifths chance for choosing no left handers. The same process will be done to see if one left hander is chosen, and two left handers are chosen. Once all of those calculations are completed, the probability for choosing one of each is four-fifths chance and the probability of selecting two left handers is a one-tenth chance.

A baseball player throwing a baseball

Description automatically generated with medium confidencePitchers are not known for their batting statistics; however, they are more known for their pitching prowess. On the statistics sheet is an AVG tab that is on both the outfielder’s side and the pitcher’s side. The one that will be focused on in this scenario is the pitching side. This stat refers to the percentage of batters that hit and get on base from the pitches that they throw out. Good pitchers will typically have this number closer to the low .200 range or even below that range. Take a look at the stats for one of the starting pitchers Aaron Nola. His pitching average is .187 which his exceptionally good for a starting pitcher. Now let us say that Nola will face fifteen batters within three innings. Out of those fifteen batters that he faces, what is the probability that four of those players will score a hit from Nola’s pitches? This can be solved using binomial distribution. The formula for binomial distribution goes as follows: This means that the combination of the sample size and the amount of successes that are requested in the problem are multiplied by the rate of success raised to the amount of successes. Once that is finished, the number is multiplied to the rate of failure raised to the power of the sample size minus number of successes. In our case, the sample size or *n* is fifteen and the number of successes that is being looked for or *y* is four. The formula will be set up like so: . Once this is calculated out, the probability for four batters to hit of off Aaron Nola in three innings is 24.8%. That is incredibly low for the randomness that comes with pitching and the complexity of throwing different pitch types as well.

Now let us change the problem slightly here. What if the problem now wants to know the chance of a batter getting a hit on the fourth batter from the fifteen batters? All of the previous values will be used like the rate of success and failure just used in a separate way. This way is known as the geometric distribution. What geometric distribution describes is the probability of success at a given number of attempts. This is different from binomial distribution that looks for the percentage of total successes and not juts one at a particular place. The function for a geometric distribution looks like so: . This will take the rate of failure and raise it to the power if the attempt number minus one. Afterwards, that value will be multiplied by the rate of success in order to get the rate of success for that particular trial. For this scenario, the formula will be laid out like this: . When doing the calculation on the left, the odds that a person will hit off of Aaron Nola on the fourth attempt at doing so is a measly 10.04%.

There are days where the Phillies light it up and crush some home runs left and right. Some games even had rallies from a deficit of four runs and they came back all within one inning. Let us say that we have a scenario where the Phillies need a comeback at the bottom of the eighth inning. Out of all ten players in the roster that can bat, five are estimated to get a home run this inning. What is the probability that out of those estimated five players, two can score a home run and take the lead in the bottom of the eighth? In order to solve a probability like this, we can use something called hypergeometric distribution. This distribution is used when comparing a sample size to a general population size. Both population and sample size will have their own failure and success rates. The formula for hypergeometric distribution looks like this: . This means that the combination of the combination of the number of successes from the sample *r* and the random variable *y*. that result is multiplied by the combination Andrew McCutchen running to home plate

of the population size *N* subtracted by *r* and the sample size *n* subtracted by *y*. Once the product is found, it is divided by the combination of the population size and the sample size. For our scenario, the number of successes from the sample is going to be five and the random variable is two. That random variable is the amount of people that the problem estimates could get a home run. The total population is ten in this case and the sample size is five. With all of that in mind the problem will be set up like so: . When running though the calculation with a hypergeometric distribution in mind, the probability of two of the five estimated batters hitting a home run is 3.9%. Looks like its going to take more than just homeruns to win the game there.

One final scenario that will be discussed is then number of strikeouts in a given series. In baseball, one team will typically play another for three to four games in a row which is called a series. In this scenario, the average amount of strike outs for one series is observed. The strikeouts are uniformly distributed between thirteen and eighteen strikeouts per series. Find the probability of a series having above fifteen strikeouts and below sixteen strikeouts. First let us break up the problem into two parts. The first one being finding the probability of having above fifteen strikeouts in a series. There are two ways to do this problem in solving the probability of uniform distribution. One would be to set up an equation like so . The value of *d* and the value of *c* are the range of what the problem asks, and the values of *a* and *b* are the range values of the entire problem. For this first part of the problem, the equation would look like this which after simplification would look like this . That means that there is a 60% chance that a series would have above fifteen strikeouts. For the other part of the question, we will use another method to solve this. It is quite simple and all it has to do is use a number line. On both ends of the number line, mark the range given by the problem on both ends. Afterwards, mark all of the points in the line by whole numbers. Once all of that is complete, just mark with an *x* or a check mark if the problem will go through those values and find the percentage of check marks to *x*’s. For this problem, the line will have a 13 and an 18 on the ends with all of the values in between marked as well. This part states that they are looking for the probability of a series having below sixteen strikeouts. So, all values that are below sixteen will have a check mark on it while everything above will have an *x*. After drawing it out, it would be the same probability as the previous part with a 60% chance of occurring.

Throughout this report, we looked through the data set of the 2021 Philadelphia Phillies to extract some key information. Finding the average doubles, triples, and home runs hit allowed us to see how the team performed for the previous year and to make improvements. Different game time scenarios that could actually happen like a late game comeback with a certain group of players were examined. The race for the best hitter on the team to receive a hefty bonus in the player’s contract. All of these scenarios and more can be solved using these formulas and methods described in this report. Statistics are everywhere, all you have to do is look around and infer.

# Statistics Sheet





# Works Cited

Adigard, E. (2008, May). *Mathamatical Statistics with Applications.* Belmont: Cengage Learning. Retrieved from unodc: https://www.unodc.org/e4j/en/organized-crime/module-9/key-issues/adversarial-vs-inquisitorial-legal-systems.html#:~:text=Common%20law%20countries%20use%20an,legal%20rules%20criminal%20procedure%20followed.

MLB Advanced Media. (2022). *Philadelphia Phillies Stats*. Retrieved from mlb: https://www.mlb.com/phillies/stats/