Colin Glory

Professor Hoy

Probability and Applied Statistics

MLB Data Set Test/Analysis

**Disclaimer**

In the MLB season of 2022, there were a total of 987 active players. The 987 players do include players when they are traded from team to team, so some players may have multiple rows of data. These specifics cases will be counted as another player.

**Chapter 2 Questions**

**Exercise 1:** The 2022 MLB season consisted of 986 players. There were 48 players with a batting average of .300 or higher, 92 players had 500+ at bats, and 9 players with over 500 at bats with a .300+ batting average . Find the number of players that were:

1. Had a .300+ batting average, 500+ at bats, or both
2. Had a .300+ batting average with less than 500 at bats

**Exercise 2:** Let A be Pete Alonso getting a hit, and let B be Jose Altuve getting a hit. Based on the table below, find:

Table

Description automatically generated

P(A): **162/597 = 0.271**

P(B): **158/527 = 0.299**

P(A’): **1-0.271 = 0.729**

P(B’): **1 - 0.299 = 0.701**

P(A ∩ B): **0.0810**

P(A ∪ B): **0.488**

P(A|B): **This would end up equaling P(A), which is 0.271, since the two events are both independent**

P(B|A): **This would end up equaling P(B), which is 0.299, since the two events are both independent**

**Exercise 3:**  (Combination) Pat is tasked with selecting any 15 batting averages from random. How many total combinations are possible with selecting 15 random averages?

**Answer: 987 C 15**

**Exercise 4:** An analyst has five seemingly identical skilled baseball players available for selection. Unknown to her, two of the five have a batting average below .150. A particular order calls for two of the players and is filled by randomly selecting two of the five that are

available.

a. List the sample space for this experiment.

**E1=(A1, A2) E2=(A1,A3) E3=(A1,B1) E4=(A1,B2) E5=(A2,A3) E6=(A2,B1) E7=(A2, B2) E8=(A3,B1) E9=(A3,B2) E10=(B1,B2)**

**S={E1,E2……E10}**

b. Let A denote the event that the selection is filled with two above .150 ba players.

List the sample points in A: **{E1,E2,E5}**

c. Assign probabilities to the simple events in such a way that the information

about the experiment is used and the axioms in Definition 2.6 are met.

**1/10**

d. Find the probability of event A.

**P(A)=P(E1)+P(E2)+P(E5) = 3/10**

**Exercise 5:** There were a total of 987 players in the MLB in 2022. Using the data set find,

1. The average of the league’s batting average
2. The standard deviation of all of the batting averages
3. The variance of the batting averages

**Exercise 6:** (Sample Points Probability) Consider an experiment that consists of recording the batting percentage for each of 20 randomly selected players. Assuming that there are only 1001 possible distinct batting averages, find the number of points in the sample space S for this experiment. If we assume that each of the possible sets of averages is equiprobable, what is the

probability that each person in the 20 has a different batting average?

n = 1001\*1000\*999…..\*982

N = (1001)^20

n/N = 0.826

**Exercise 7:** The names of 34 players of each league are to be randomly drawn, without replacement, from a bowl containing the names of 987 players of the MLB for the All Star Game. How many sample points are associated with this experiment?

**Exercise 8:** A league analyst it tallying up the total league at bats to determine the average season at bats for the league. Using the dataset, find the league average at-bats.

**Exercise 9:** Ronald Acuna Jr had 29 stolen bases in 2022. Let’s say in one game, Ronald Acuna got on first base 2 times. Let S denote to Ronald Acuna stealing a base, Let N denote to him not stealing a base. Let S denote to the possible outcomes:

Let A denote the subset of possibilities containing no stolen bases; B, the subset containing two

steals; and C, the subset containing at least one stolen base. List the elements of

**A: {NN}**

**B: {SS}**

**C: {NS, SN, SS}**

**Chapter 3**

**Exercise 1:** Ronald Acuna Jr has a batting average of 0.266. What is the probability of him getting at least one hit in 4 at bats?

**= 1 - (1-.266)^4 = .710**

**Exercise 2:** Adam Duvall had a 0.213 batting average in 2022. Let’s say in one game, he gets 4 at bats.

A. What is the probability he will get 1 hit in the game?

1. What is the probability he will get 2 hits in the game?
2. What is the probability he will get 3 hits in the game?
3. What is the probability he will get 4 hits in the game?

**Exercise 3:** Using results from the previous problem, we get the table below.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| X | 1 | 2 | 3 | 4 |
| P(X) | 0.415 | 0.169 | 0.0304 | 0.00206 |

(A) Compute and interpret the mean of the random variable X

(B) Compute the standard deviation of the random variable

**Exercise 4:** Matt Duffy wants to get a hit in 10 at bats. The probability that he is successful on a given trial is .25.

a What is the probability that the third at bat is the first to yield a hit?

b If Matt can get at most 10 at bats, what is the probability that he will fail to get a hit?

**Exercise 5:**  From a group of 20 MLB Players, 10 are randomly selected for All Star infield players(including reserves). What is the probability that the 10 selected include all the 5 best players in the group of 20?

**Exercise 6:** There were a total of 5714 home runs in the 2022 MLB season. The mean number of homeruns were 5.78, with a standard deviation of 8.08 homeruns. Apply Chebyshev’s Theorem to the data using k=2. Interpret the results.

According to Chebyshev’s Theorem,

K is equal to 2.

Proportion Value = 75%

So, if k= 2

75% Chebyshev’s interval are:

**Exercise 7:** Let Y be a random variable with p(y) given in the accompanying table.

1: Shohei Ohtani OBS 2:Matt Olson OBS 3: Marcell Ozuna OBS 4: Rafael Ortega OBS

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **y** | 1 | **2** | **3** | **4** |
| **p(y)** | **.356** | **.325** | **.274** | **.331** |

= 3.152

= 0.692

= 8.418

=

**Results**

This was an interesting way of doing a project assignment. Finding a data set to create and answer questions based on the data. I decided to choose a data set involving the 2022 season in the MLB. This data set was interesting to work with due to being able to see and analyze every single player’s statistics for that season.

The section of data I decided to mess with was the batting average. It was shocking to see that the league’s average for batting averages was 0.197. A 0.197 batting average is not the greatest. A batting average of .250 would be an acceptable and more accurate average in the MLB. A 0.197 would be for a player who’d typically bat lower in the lineup. The highest batting average according to the data set is 0.750 by Chadwick Tromp, but he only had 4 at bats. It’s the same thing for the next 16 players, so we can remove those to find the batting average leader. The leader for that year is Jeff McNeil with a batting average of 0.326, which is a fantastic average.

It was interesting to be able to figure out the probability of a player getting x amount of hits per game based on his batting average. For example, exercise 2 deals with finding the probability of Adam Duvall getting a hit based on his batting average and the amount of at bats in the game. He had an average of 0.213 in 2022. I was able to figure out the probability of him getting 4 hits in a game with 4 at bats, 0.00206, which is incredibly low odds of happening. This same process can be applied to other areas of the dataset, a fairly simple process.

Major League Baseball batting statistics are a crucial aspect of the sport, providing insights into player and team performances. These statistics enable fans, players, coaches, and analysts to track individual and team achievements, compare different players, and assess player value. The most common MLB batting stats include batting average, on-base percentage, slugging percentage, and on-base plus slugging. These stats provide valuable insights into player performances, such as a player's ability to hit for power or to get on base, and help teams make informed decisions about player acquisition, lineup construction, and player development.

Batting statistics play a critical role in evaluating player performance and determining player value. Teams use these statistics to assess the contributions of individual players to their team's success and make decisions on player salaries and contracts. For example, a player with a high batting average or on-base percentage is considered a valuable asset to a team, while a player with a low batting average or on-base percentage may be seen as less valuable. Batting stats also help identify players who are consistent performers or who are trending in the right direction, making it easier for teams to make informed decisions on player acquisition and development.

Another reason why MLB batting statistics are important is that they enable fans to track the performances of their favorite players and teams. Fans can use these statistics to compare different players and evaluate their contributions to their team's success. In addition, the use of batting statistics has led to an increased interest in the game among casual fans, as they provide a deeper understanding of the sport and its players.

Finally, MLB batting statistics are crucial for players and coaches to evaluate their own performances and make adjustments to their game. By analyzing their own batting statistics, players can identify areas where they need to improve and adjust their approach to the game. Coaches can also use these statistics to help players improve their technique and develop their skills. Using a dataset of these statistics can enable a person to predict the probabilities of certain players getting X amount of hits in a game, as well as many other aspects of hitting. Overall, MLB batting statistics play a vital role in the sport, providing valuable insights into player and team performances and helping teams make informed decisions about player acquisition, lineup construction, and player development.

Works Cited

Wackerly, D.D., Mendenhall, W. and Scheaffer, R.L. (2008) Mathematical Statistics with Applications. 7th Edition, Thomson Learning, Inc., USA.

https://www.kaggle.com/datasets/vivovinco/2022-mlb-player-stats