Part 1: Data

Anna, Riley

2/7/2022

Overview of the Dataset

This data was collected almost entirely from baseballsavant.com. The categorical variables were collected from baseballreference.com and mlb.com. Our sample population is the 132 Major League Baseball players who had more than 502 plate appearances during the 2021 season, enough plate appearances to qualify for end-of-year awards. Batters in the sample have between 505 and 724 plate appearances. We chose qualified players out of a desire to avoid treating players with far fewer plate appearances (smaller sample sizes) with equal statistical significance. Each row/observational unit represents one batter.

The columns are the variables, 4 of which are independent categorical variables. These are handedness (what side of the plate the batter bats from), position (what field position the batter plays the most innings at), mode of acquisition (high school draftee, college draftee, or international free agent signing), and the division they play in. The other 41 variables are quantitative. Some quantitative variables are/have been made discrete, such as age or number of home runs. The quantitative variables of most interest to us are:

b_hr_rate: the rate at which the batter hits home runs. Home runs/plate appearances.

b_k_percent: the rate at which the batter strikes out. Strikeouts/plate appearances.

b_bb_percent: the rate at which the batter walks. Walks/plate appearances.

batting_avg: the rate at which the batter gets base hits. Hits/at bats.

isolated_power: where a single is worth 1 total base, a double is worth 2, and so on, $ISO = (totalbases - singles)/atbats$

exit_velocity_avg: the average velocity of the hitter's batted balls.

launch_angle_avg: the average vertical angle at which the ball leaves the hitter's bat, where 0 degrees is parallel to the ground.

sweet_spot_percent: the percentage of the hitter's batted balls that are hit between 8-32° launch angle.

z_swing_percent: the percentage of pitches in the strike zone that the batter swings at.

oz_swing_percent: the percentage of pitches outside of the strike zone that the batter swings at.

whiff_percent: the percentage of pitches at which the batter swings and misses entirely.

pull_percent: the percentage of a hitter's batted balls that are hit to the same side of the field as the batter's box they bat from (left/right).

groundballs_percent: the percentage of a batter's batted balls that are hit on the ground (less than 10° launch angle).

flyballs_percent: the percentage of a batter's batted balls that are hit in the air, between 25-50° launch angle.

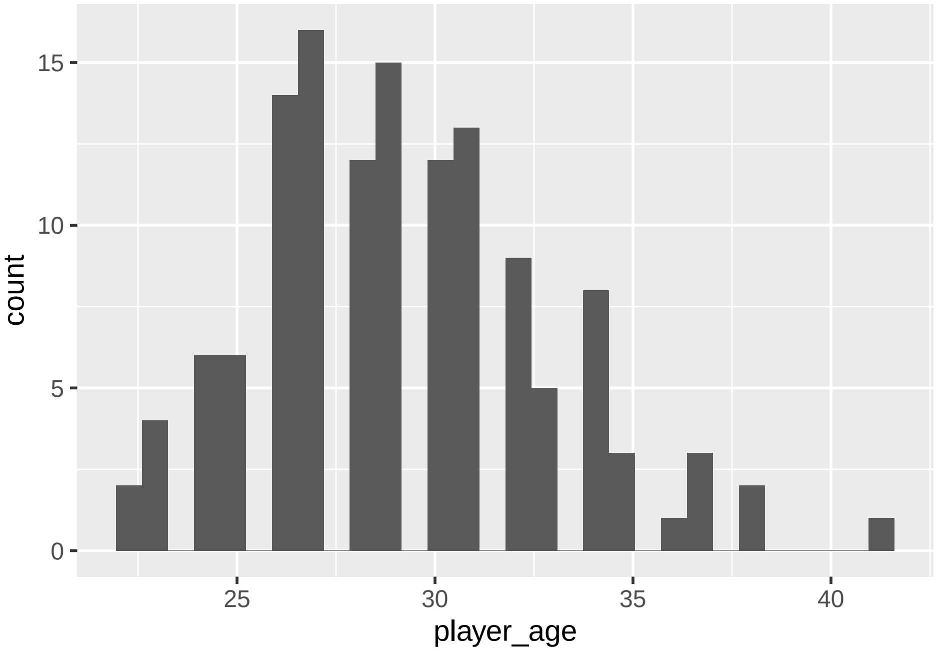
linedrives_percent: the percentage of a batter's batted balls that are hit between 10-25° launch angle.

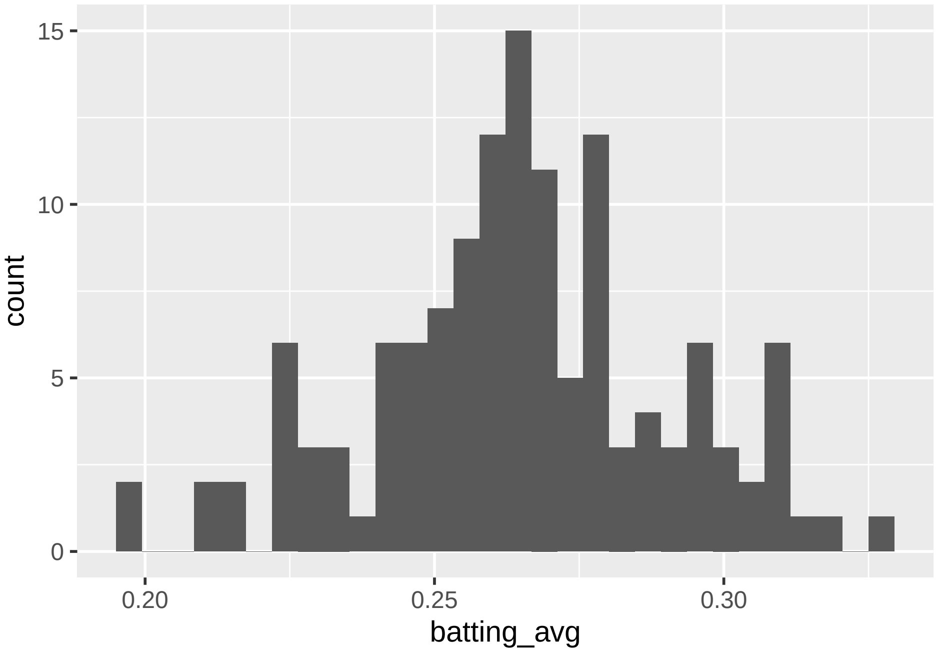
Summary Statistics

Perhaps the most important summarizing statistic that can be found here is the standard deviation in the total number of plate appearances (PA). One standard deviation is equal to 56.2 about a mean of 597. This is significant as it pertains to counting stats (as opposed to averages/rate stats). For example, if we were looking to compare players on their ability to hit home runs, simply comparing home run totals would not be the best way because of the variance in number of PAs. It would be better to use home run rate, HR/PA. This variance in sample size matters far less for averages/rate stats because they are adjusted for number of PAs and averages/rates tend to approach their "real" values in large samples (502 PAs being a pretty large sample). In our future analyses of this data set, we will confine ourselves to averages/rate stats.

We can gain further insight by comparing some of the means here to the overall 2021 MLB means. For example, the league-wide slash line (batting average/on base percentage/slugging percentage) in 2021 was .244/.317/.411. (see: https://www.baseball-reference.com/leagues/majors/2021.shtml). In this sample, the mean slash line is .265/.339/.457, which is better. This makes sense, as better players tend to be given more plate appearances, while many others do not receive enough to qualify. While some means in this sample may vary some from the greater population of MLB players, it may be a necessary sacrifice. In order to conduct linear regression tests that give us the most accurate possible view of the relationships between these variables, each player needs to have a sufficiently high number of plate appearances. This way, we can avoid small sample noise. Lowering the plate appearance requirement for this data set would make this sample more similar to the average player, but there would still be difference in the same direction. To be able to compare the two, we have have to sample almost the entire population. We will say that this sample is taken from the population of MLB players with roughly a full season's worth of plate appearances (502+).

Graphical Analysis





```
## [1] 0.265
```

- Graphical displays with adequate explanation / interpretation (These should effectively summarize your data and point out any interesting features.
- quantitative vs quantitative - categories in colors
- no linear models

Discussion

- discuss the limitations of describing a larger population.