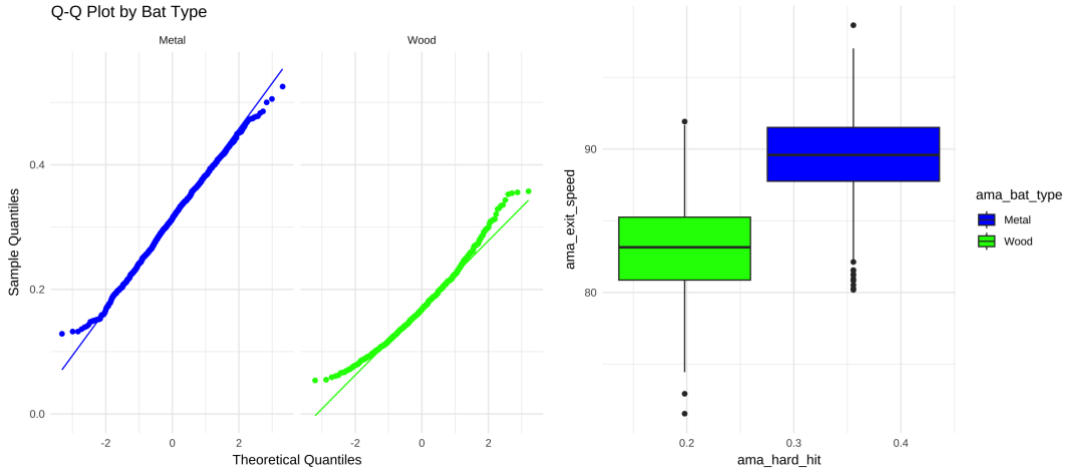
The first step in any statistical analysis is cleaning the data. I cleaned the data by eliminating players with minimal amateur plate appearances (PA > 50). I also eliminated players with no hard-hit percentage for their professional experience and no exit data for their amateur careers. This cleaned data gives us our sample from the population. Following the filtering and cleaning of our data, we must run some diagnostic tests to determine what test to use to determine whether there is a statistical difference between bat types and whether we can find a correlation between amateur data and professional data.

We must ensure that our data fits the assumptions required for a two-sample t-test. These assumptions include normality and equal variances for the data. To determine normality, we visualize the data on a Q-Q Plot that helps us determine normality. A boxplot and Brown-Forsyth-Levene test help us determine if variances are equal or not.



We can see on this Q-Q Plot that our sample data is normal; however, the wood bat bata does have some minor curvature at the ends and could indicate non-normality. A Brown-Forsyth-Levene (BFL) test is used to test for equal variances. A BFL test run on the data in R gives us the p-value of 5.466e-12, which is < 0.05 (significance level), and we reject our H_0 that the variances are equal. Given the results of our BFL test and the visualization of the Q-Q Plot, it would be best to use a nonparametric test rather than a two-sample t-test. The nonparametric test that I decided to use is the Wilcoxon Rank Sum test because it is one that I am most familiar with.

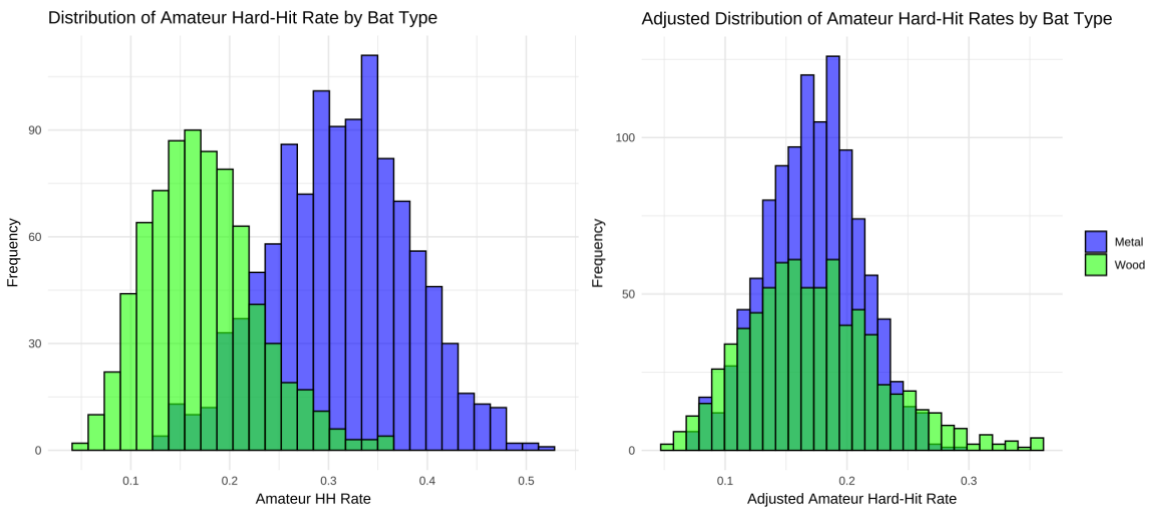
The hypotheses used for the Wilcoxon Rank Sum test:

H_0 : Two Populations are Equal

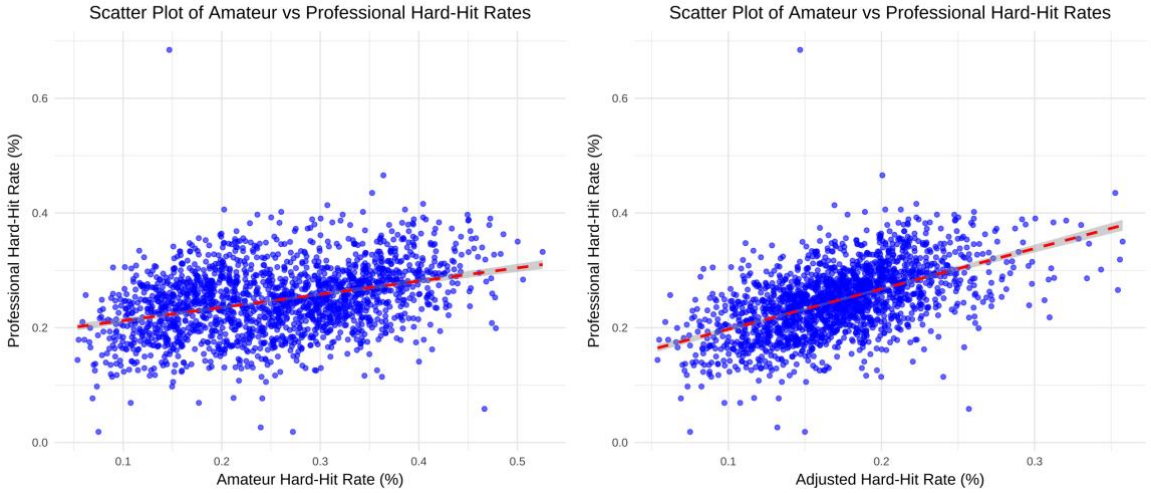
H_a : Two Populations are not Equal

Using R to run our test, we get a p-value of 2.2e-16 < 0.05; therefore, we reject our H_0 and conclude that the two populations are unequal, or the hard hit % for wood and metal differ.

Given this information, we must determine an appropriate adjustment factor for our data. We can use the ratio of mean HH% (wood) to mean HH% (metal) as our adjustment factor. This ratio gives us 0.17218/0.3125 or ~0.5509. We apply this adjustment factor to all amateurs with metal bat data by multiplying their HH% by 0.552 to obtain their "adjusted_hard_hit." The non-adjusted and adjusted data can be visualized with histograms.



Following this adjustment to our data, we can run a correlation analysis to determine if the relationship was strengthened between amateur and professional HH%. Using the Pearson correlation coefficient, we can find the correlation coefficients for the original and adjusted data. The data can also be plotted on a scatter plot to visualize the correlation change.



Before: 0.3631 After: 0.5426

By assessing the visualizations and the change in correlations, we can determine that our adjustment factor effectively strengthened the relationship between pro and amateur hard-hit rates.