Wind Speed vs. Total Errors Committed Project Report

**Research Scenario:**

For this project, I was interested in examining how wind speed affects the number of errors committed during MLB games. For this purpose, I took a random sample of 500 games from a dataset containing information from baseball-reference.com on each MLB game played during the 2016 season. From this sample, I examined the total errors committed during games where reported wind speed was between 0 and 10 mph and where reported wind speed was between 11 and 25 mph. I then compared these two variables via a two-sided t-test to see if any evidence indicated that the number of total errors was greater between 11 and 25 mph than for wind speeds between 0 and 10 mph and performed this test at the alpha = 0.10 level of significance.

**Description of Data Set:**

The two columns utilized in my analysis were the wind\_speed and total\_errors columns. The wind\_speed column listed the wind speed recorded at game time in mph for each game and the total\_errors column reported the total number of errors committed by both teams during each game. Before performing any operations on the data, I first minimized the number of rows in the original dataset from 2,463 to 500 games by taking a random sample of 500 games from the original 2,463 and saving this newly created data frame under sample\_data. This was done in order to simplify the analysis process by using a portion of the number of games played to make an educated guess as to whether higher wind speeds had a greater impact on total errors. I then created the total\_errors column by adding together the data contained in the away\_team\_errors and home\_team\_errors columns for each game and added this to sample\_data. This column allowed for an easier and more uniform comparison of errors to wind speed. I then removed all other unused columns in sample\_data to reduce unnecessary clutter in the data set. I neglected to remove any outliers as the number of errors committed and the wind speeds reported were observed values that occurred during games. Also, the goal of this research was to measure whether higher wind speeds impacted total errors more than lower wind speeds, and since you can’t have negative wind speeds or negative errors, it made sense to include high wind speed observations. As a final step, I checked to make sure that each of the assumptions for a two-sample t-test were met. The assumption of same measurement was met as both as both variables of interest were measured according to number of errors committed. Similarly, the assumption of similar distributions was met as the histograms of both variables closely resembled each other. Finally, the assumption of independence was met as both samples were randomly selected from different populations and did not have any influence on each other since errors committed during games where wind speed was between 0-10 mph and can’t also be committed during games with recorded wind speeds between 11-25 mph.

Link to main dataset source: <https://www.kaggle.com/datasets/cyaris/2016-mlb-season>

**Research Question:**

The research question for this project was whether wind speeds between 11 and 25 mph had a greater impact on the total number of errors committed than wind speeds between 0 and 10 mph during 2016 MLB games. To examine this further, I utilized a two-sided t-test to examine whether the mean number of errors committed between wind speeds of 0 to 10 mph was less than the mean number of errors committed between wind speeds of 11 to 25 mph and a boxplot showcasing the distribution of wind speeds for each number of total errors committed.

**Conclusion:**

The t-test results indicate that we don’t have significant evidence to conclude that the number of total errors committed between wind speeds of 0-10 mph is less than the number of total errors committed between 11-25 mph. Wind speed does not appear to have a huge effect on number of errors committed, but more research needs to be performed. The plot below details reported wind speeds for each additional error committed throughout the 500-game sample.

Chart, box and whisker chart

Description automatically generated

This supports the t-test’s conclusion as the distribution of wind speeds for each increase in the number of errors generally remains consistent. The exception is games where 6 total errors were committed, as the distribution of wind speeds is slightly higher than the rest at that level. However, this is more a result of the limited number of games where 6 errors occurred rather than wind speed directly effecting errors.