M119 Exam Chapters 3-5 Do-At-Home Part
P Staley Baseball Stats Analysis

Batting average is a measure of how well a team is hitting. Earned run average is a measure of how well a team is pitching (There is a table of baseball terms/abbreviations on the DAH Data page). Each student has been assigned a year to use for the following analysis of baseball statistics (see www.staley-classes.org website under the stats column Exam 2 Do-At-Home Data).

1. **Compare with Given Distribution**. If the long term density curve for major league team-batting-averages is a Normal distribution with mean .267 and standard deviation .01 then

a. What batting average would we expect San Diego needs to be in the (long term) top 40% of major league teams? (Based on our hypothetical long term Normal distribution)
b. What team batting average would you expect to be Q1? Q3? (for the long term hypothetical density curve).
c. Find the team-batting-average IQR for your assigned year and compare it with the long term hypothetical density curve. What can you say about the spread of team-batting-averages for your year vs the spread of the long term hypothetical distribution?

2-5 below refer to your assigned year:

2. **Collect Data**. Make a table with the following columns: team, batting average, earned run average, and win percentage. You will need to use more than one column to compute the win percentage. Print out the table along with your explanation of how you computed the win percentage. Be sure to check your data table for accuracy.

3. **Standardized Scores**. Use your data table to compute means and standard deviations and then use those to compute z-scores for team-batting-averages, team-earned-run-averages and win percentages. Add the z-scores as columns to your table and print out the result.

4. **Predicting Wins from Batting Average**. Using the data for your year, make a scatterplot with batting average as the explanatory variable and win percentage as the response variable. Print out your scatterplot.

a. What is the correlation of batting average to win percentage?
b. Find the linear regression of win percentage to batting average.
c. The variation in win percentage from team to team can be the result of many things—opponents mistakes, luck, umpire decisions, quality of pitching, quality of hitting, etcetera. How much of the variation in win percentage can be explained by differences in hitting ability as measured by the team-batting-averages?

5. **Predicting Wins from ERA**. Using the data for your year, make a scatterplot with team-earned-run-average as the explanatory variable and win percentage as the response variable. Print out your scatterplot

a. What is the correlation of team-earned-run-average to win percentage?
b. Find the linear regression of win percentage to team-earned-run-average.
c. The variation in win percentage from team to team can be the result of many things—opponents mistakes, luck, umpire decisions, quality of pitching, quality of hitting, etcetera. How much of the variation in win percentage can be explained by differences in pitching ability as measured by the team-earned-run-averages?

6. **Observations** Write a paragraph describing in layman's terms what can be gleened from the data you examined. Were there any surprises? Any outliers? Did you draw any conclusions from comparing the results of #4 and #5. What other sort of things could be done with this data to better understand what produces baseball wins? What other variable would be interesting to add to these tables/analyses?

A Note about Scoring. Here is a list of things that students in past semesters have done that have adversely affected their score on this DAH exam:

a. They didn't label/title their graphs/tables.
b. They didn't put the year of interest in the graph/table title.
c. They didn't capitalize the first letter of each word in the graph/table title.
d. They did not specify what means and standard deviations they used for #3.
e. They had misspellings.
f. They included extraneous columns in the data tables for item #2. Remember this is about making a clear presentation to your customer. Filling your report with random cr*p will make your customer think that you don't know what you are doing.
g. They did not print out a table as specified in #2. (Then they tried to argue for counting the table in #3 as the table in #2).
h. Their tables/graphs unnecessarily lapped over to two pages.
i. Two page tables did not have column/row headers on the second page. The customer should be able to look an any number of any report and easily understand what that number is about.
j. Students turned in a picture and thought that was the linear regression.
k. Students made the answers hard to find.
l. They did not check their data to make sure that each team had the correct batting average, earned run average, and win percentage. This is an amazingly common error no matter how many cautions I put out. Being careful is an important part of this class.
m. They mislabeled wins percentage as wins or vica versa.
n. They used "batting-average-allowed" which is a pitching statistic instead of team-battng-average.
o. They used "hits against" (a pitching statistic) as a team-hitting statistic.
p. They didn't label the axes of their graphs.
q. They refused to use the given density curve for #1.
r. They did not scale the scatter plots so that the points filled the space available.