UNITED STATES MILITARY ACADEMY

FINAL PROJECT

MA388 SABERMETRICS

SECTION D1

LTC MICHAEL POWELL

BY

CADET AIDEN J. BOEHM ’24, CO E2
CADET ETHAN FUHRMAN ’24, CO I4

WEST POINT, NEW YORK

10 MAY 2024

AB EF MY DOCUMENT IDENTIFIES ALL SOURCES USED AND ASSISTANCE RECEIVED IN COMPLETING THIS ASSIGNMENT.

____ I DID NOT USE ANY SOURCES OR ASSISTANCE REQUIRING DOCUMENTATION IN COMPLETING THIS ASSIGNMENT.

SIGNATURE: ____

Works Cited (Along with MA206 Course Guide):

Lane, David. *Introduction to Statistics*. 2003, https://onlinestatbook.com/2/index.html.

Marchi, Max, et al. "Analyzing Baseball Data With R." *Chapman and Hall/CRC eBooks*, 2018, https://doi.org/10.1201/9781351107099.

SeanLahman.com. www.seanlahman.com.

"Standard Error, N." Oxford English Dictionary, Oxford UP, March 2024, https://doi.org/10.1093/OED/6084596956.

Are Home Runs the Key to a Championship?

Aiden Boehm and Ethan Fuhrman

5/5/2024

Learning Goals

1. Understand a standard error and standard deviation and how they affect an estimate's interpretation and significance.
2. Understand how to construct different confidence intervals such as 67%, 95%, and 99% intervals and how they inform are thinking about home runs and team success in baseball.
3. Understand how to interpret a confidence interval and how it informs statistical significance in terms of linear regression and two-sample t-test.
4. Interpret statistical difference in baseball context to determine if World Series winning teams, on average, hit more home runs than those that do not win the World Series.

Introduction

This module will seek to present different ways to interpret and visualize standard errors, confidence intervals, and statistical significance.

1. Standard error is defined as a measure of the statistical accuracy of an estimate, equal to the standard deviation of the theoretical distribution of a large population of such estimates (Oxford English Dictionary). The standard error is then used to create confidence intervals.
2. A confidence interval is constructed using the standard error with the formula of the measured statistic plus/minus the multiplier associated with that confidence interval multiplied by the standard error of the statistic. It then asserts the confidence level by determining that there is a 95% chance the true value of the statistic is within that range outlined by the confidence interval.
3. A two-sample t-test analyzes the difference between two statistics to determine if it is significant based on the t-distribution. Validity conditions must also be met for this test to take place such as the data being distributed normally.

In order to understand these methods we will investigate the effects of home runs on a team's success. A home run is a hit in baseball which the batter makes it around all the bases and scores a run with no errors ocurring in the field. More specifically we will look at if a team that wins the World Series hits more home runs than their peers in a season. We will investigate this difference and determine if it is statistically significant and if so at what confidence levels.

Data

To begin the exercise, we will first construct the data frame which we are about to examine. Below you will find the code to do this. First, we create a new data frame called 'spec_dat' and assign the 'Teams' data set from the Lahman Baseball Database to the data frame. The Lahman Baseball Database is a compilation of data on pitching, hitting, and fielding performance – as well as other tables – from 1871 through 2022. We chose to use only the years 1980 – present (excluding 1994 due to the World Series not being played because of the lockout) because modern baseball is significantly different in terms of skill and playstyle than the early years of baseball. The data frame can be found at this link: http://www.seanlahman.com/, but it is also a package in R which was used for this project. Once it is downloaded in can be called simply by using the library command and then putting Lahman into the parentheses. From the 'Teams' data set we take the home runs and games variable. Both variables are numeric and

Methods/Instructional Content

Texts to look at and reference: MA206 Course Guide and Text at: file:///C:/Users/Aiden.Boehm/Desktop/MA206%20AY22- 1%20Course%20Guide.html#content

a. This source provided R code snippets that walked through creating confidence intervals to evaluate a data frame. It also provides some of the general information regarding the statistical definitions of the concepts listed above.

Introduction to Statistics by David Lane, Rice University at https://onlinestatbook.com/2/index.html.

a. This source provided theoretical explanations for standard error, standard deviation, confidence intervals, linear regression, and two-sample t-tests.

Analyzing Baseball Data with R (Third Edition) by Marchi, Albert, and Baumer (for information on data sets and R help) at https://beanumber.github.io/abdwr3e/.

a. This source provided numerous examples of how to investigate baseball problems though statistics using the Lahman Baseball Database. Section 2.3.1 provides explanations on basic dplyr functions to clean the data and create the sub sample desired.

Overall, a two-sample t-test will be conducted below. The validity condition of a non-skewed distribution will be shown using a histogram and confidence intervals will be constructed. The formula used to construct confidence intervals will be: $MeasuredStatistic + / - multiplier * (StandardErrorofMeasuredStatitsic)$ In this scenario the multiplier corresponds to the confidence level you are seeking to obtain. The multipliers are gathered by using the qt function in the activities below. A p-value is also calculated following the confidence interval after we state our hypothesis.

Exercises/Activities

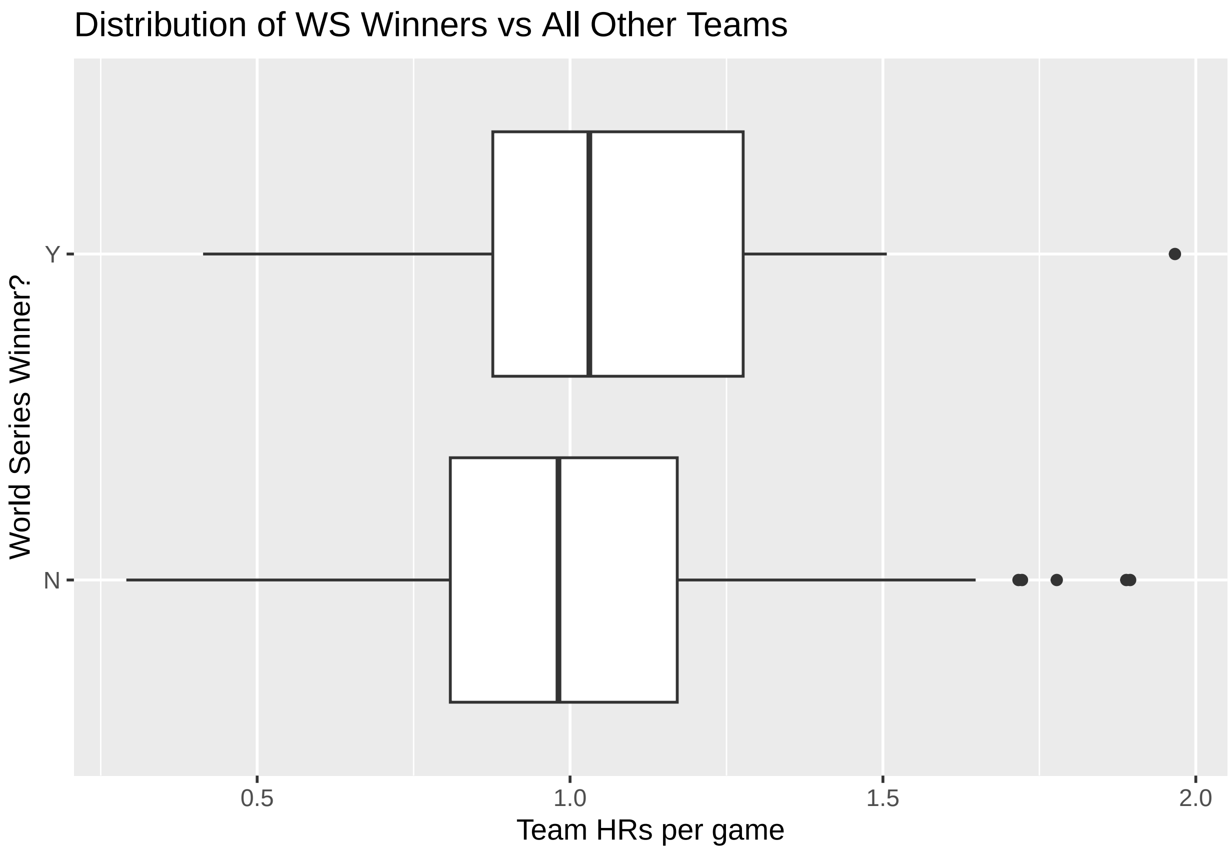
First, we filter the data set to include only years after 1980 and not include 1994 as the World Series was not played that season.

```
# Call Teams data frame
spec_dat <- Teams %>%
  # Select necessary Variables
  select(teamID, yearID, G, HR, WSWin) %>%
  # Create Home Runs Per Game Variable
  mutate(HR_per_Game = HR / G) %>%
  # Narrow down data set to only desired years
  filter(yearID >= 1980) %>%
  filter(yearID != 1994)
```

Visualize the Difference

Boxplot:

```
# Create bosplot of WS Winners and Non-Winners
spec_dat %>%
  ggplot(aes(x = HR_per_Game, y = WSWin))+
  geom_boxplot() +
  labs(title = "Distribution of WS Winners vs All Other Teams", x = "Team HRs per game",
       y = "World Series Winner?")
```



Do you think the difference in the total number of home runs per game between World Series winner and non winners is significant? If so, what could make it significant?

Answer: There does appear to be a difference from the graphs, but further testing required to determine if the difference is significant.

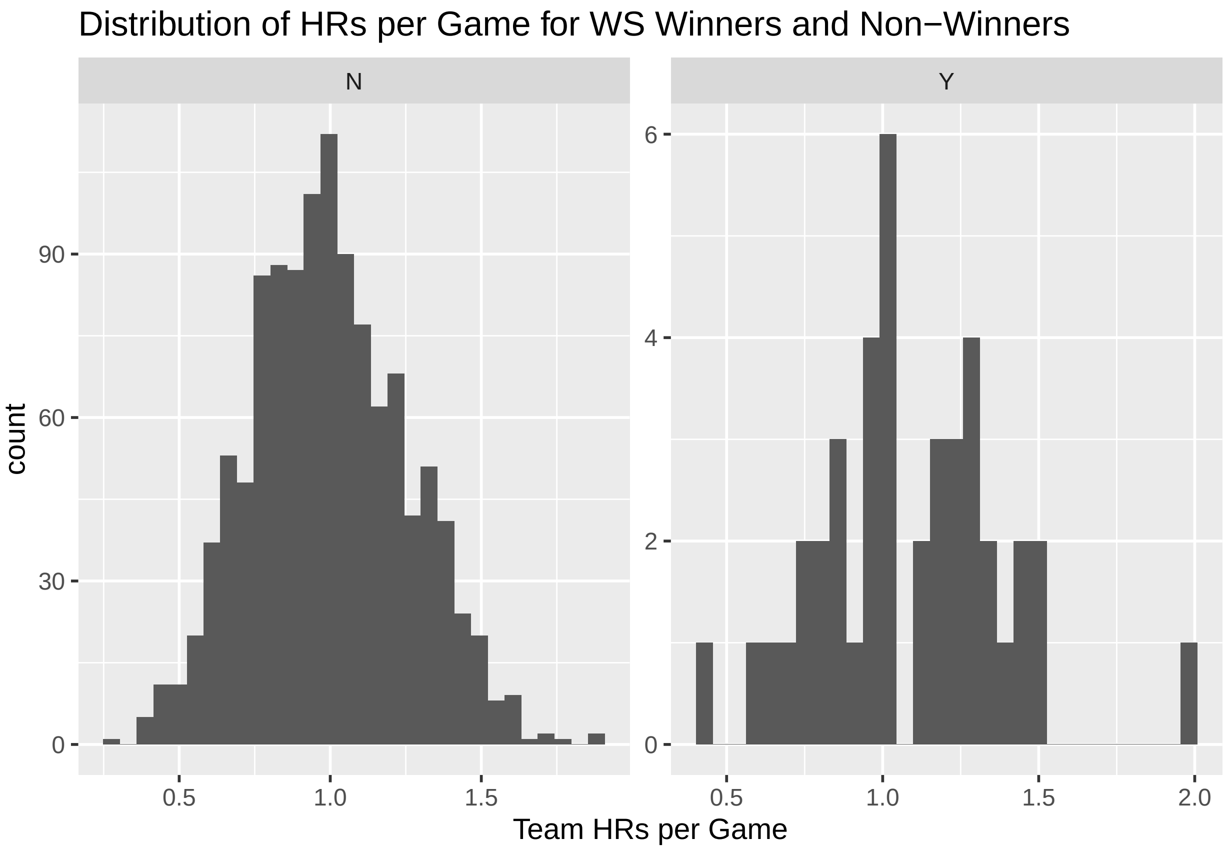
Histograms (checking validity condition for t-test):

```
# Create data set with just WS Winners
WSWin_df <- spec_dat %>%
  filter(WSWin == "Y")
# Create data set with just WS Non-Winners
NoWSWin_df <- spec_dat %>%
  filter(WSWin == "N")

WSWin <- WSWin_df$HR_per_Game
NoWSWin <- NoWSWin_df$HR_per_Game

# Create the Histogram
spec_dat %>%
  ggplot(aes(x = HR_per_Game))+
  geom_histogram()+
  facet_wrap(~ WSWin, axes = "all", scales = "free")+
  labs(title = "Distribution of HRs per Game for WS Winners and Non-Winners"
       , x = "Team HRs per Game")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Obtaining the Statistics:

Gather mean, standard deviation, and number of observations in each group (WS Winners and Non WS Winners)

```r
# Gather the necessary statistics
spec_dat %>%
  # Grouping by WS Winner and Non-Winners
  group_by(WSWin) %>%
  # xbar is the mean for the group, s is standard deviation, n is the size of the group
  summarise(xbar = mean(HR_per_Game),
            s = sd(HR_per_Game),
            n = n()) %>%
  # Present results in a table
  kable(digits = 3)
```

WSWin	xbar	s	n
N	0.995	0.255	1158
Y	1.078	0.291	42

As depicted above there are many less World Series winners than teams who did not win so there is a higher standard error as it is a smaller sample size. Next we will introduce standardized statistics and specifically the t-statistic. This statistic takes data points and scales it by a "normal" population to get a value that can be compared and and used to determine statistical significance. It uses the standard deviation to determine how far away from the mean a value is plausible simply due to random variation.

We can now calculate the standardized statistic with this given information (t-stat).

```r
# Code values in from prior findings
xbar_No = 0.995
xbar_Yes = 1.078
s_No = 0.255
s_Yes = 0.291
n_No = 1158
n_Yes = 42
# Find sample standard error
se = sqrt(s_Yes^2/n_Yes+s_No^2/n_No)
null = 0
# Find difference in means
statistic = xbar_Yes-xbar_No
# Find how different that is from the mean and divide it by the standard error
t = (statistic-null)/se
t
```

```
## [1] 1.823244
```

These calculations above show how the standard deviations for both groups are used to determine significance and meaningful differences.

Next we will set up a hypothesis test to determine a p-value associated with the t-statistic. Since we believe World Series winning teams to hit more home runs than those who did not win our alternative hypothesis would be the difference in home runs hit per game between World Series winning teams and non World Series Winning teams is greater than 0, with the null being that it is equal to 0. We can now calculate the p-value from the t-statistic.

```
# Find overall number of observations in the data set
n = n_Yes+n_No
# Find the p-value associated with the t-stat found above and
# the size of the data set
pvalue = pt(t,n-2)
1 - pvalue
```

```
## [1] 0.03425779
```

Now we can calculate the confidence intervals at the 67, 95, and 99% levels.

```
# 67% level
# Find multiplier value for 67% level
multiplier = qt(.667,n-2)
# Use confidence interval equation outlined above
CI = c(statistic - multiplier*se, statistic + multiplier*se)
CI
```

```
## [1] 0.06334528 0.10265472
```

```
# 95% level
# Find multiplier value for 95% level
multiplier = qt(.95,n-2)
# Use confidence interval equation outlined above
CI = c(statistic - multiplier*se, statistic + multiplier*se)
CI
```

```
## [1] 0.008062955 0.157937045
```

```
# 99.5% level
# Find multiplier value for 99.5% level
multiplier = qt(.995,n-2)
# Use confidence interval equation outlined above
CI = c(statistic - multiplier*se, statistic + multiplier*se)
CI
```

```
## [1] -0.03444725  0.20044725
```

These results show that there is a significant difference at the 67% and 95% confidence levels, but not at the 99.5% level. You can tell do to the fact the confidence interval at the 99.5% level contains 0 within it meaning there may be no difference whereas the other 2 intervals have both bounds greater than 0. The 95% confidence interval can be interpreted as a statistician being 95% confident that the difference in home runs hit per game by World Series winning teams on average is between 0.008 and 0.15 higher than those teams that do not win the World Series.

Two-Sample T-Test using R

These same steps can all be consolidated into 1 command using R as well as shown below.

```r
# T-test conducted using R command
t.test(WSWin, NoWSWin, alternative = "greater")
```

```
##
##  Welch Two Sample t-test
##
## data:  WSWin and NoWSWin
## t = 1.8386, df = 43.318, p-value = 0.03641
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.007193426         Inf
## sample estimates:
## mean of x mean of y
## 1.0783751 0.9945857
```

Wrap-Up Conclusion

The findings of this module show that during a 162 game season on average World Series Winning teams hit 13.5 more home runs than those teams that do not win the World Series. This difference over the course of a season is not extremely large but significant and something teams may want to keep in mind to in mind when they consider who to add to their roster. This difference is significant at the 67% and 95% confidence levels but not at the 99.5% level. So while we have strong evidence up to the 95% level, there is a small amount of doubt as it is not significant at our highest confidence level. These results were consistent regardless of the method and this module explored how sample size, standard deviation, standard error, and t-statistics are used to determine significance. There should also be a greater understanding of how to interpret a confidence interval and utilize R to calculate these different statistics. A future topic that could also explore this difference are linear regression models. These findings could also serve as a foundation to determine what the value of a home run is and what truly defines a championship team.