MLB 2017

Jensen Holm & Kyle Knapp

2022-11-07

Libraries

```
library(tidyverse)
library(boot)
```

Read in the data

```
statCast <- read.csv(paste(getwd(), "/data/statcast2017.csv", 
bbref <- read.csv(paste(getwd(), "/data/bbref2017.csv", sep = 
```

Joining the two datasets together with dplyr

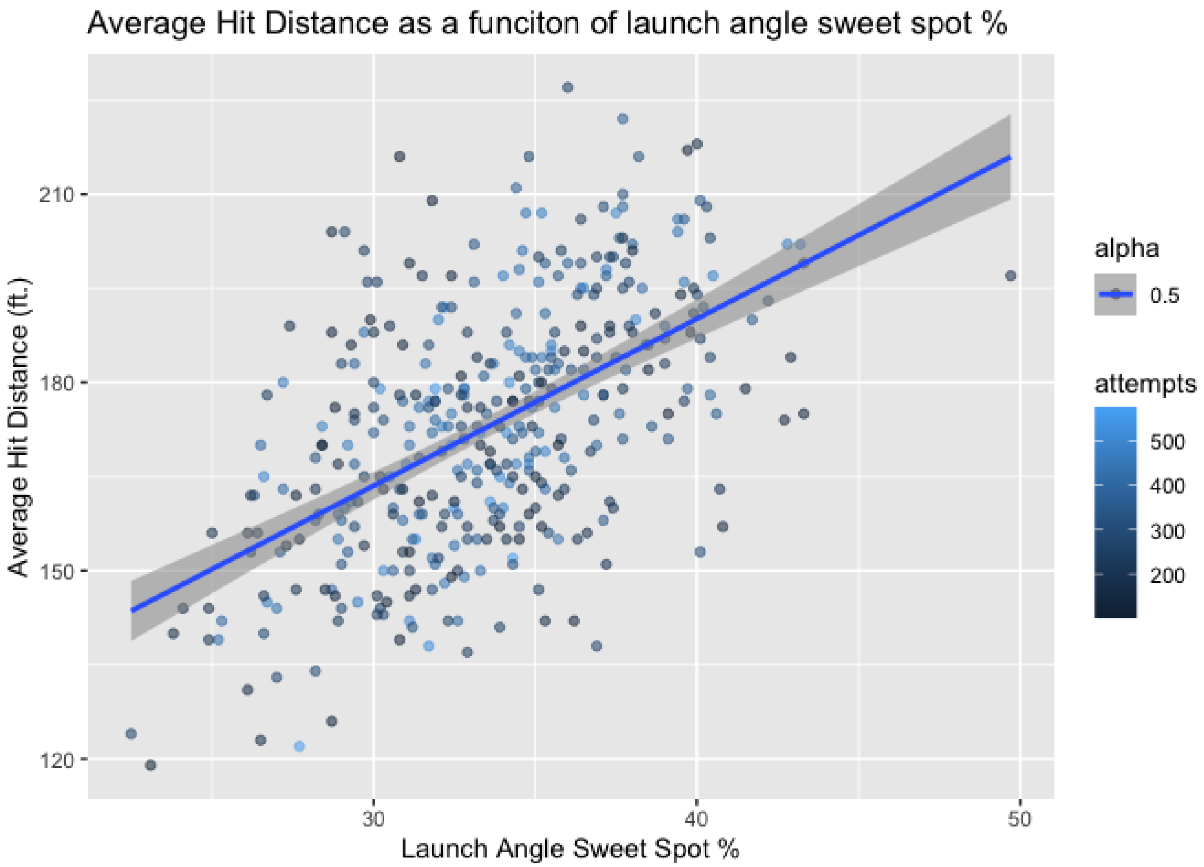
```
statCastName <- statCast %>% mutate(
  Name = paste(
    str_replace_all(
      first_name, " ", ""),
    str_replace_all(
      last_name, " ", ""),
    sep = " "))

bsbl <- left_join(statCastName, bbref, by = "Name")
```

EDA

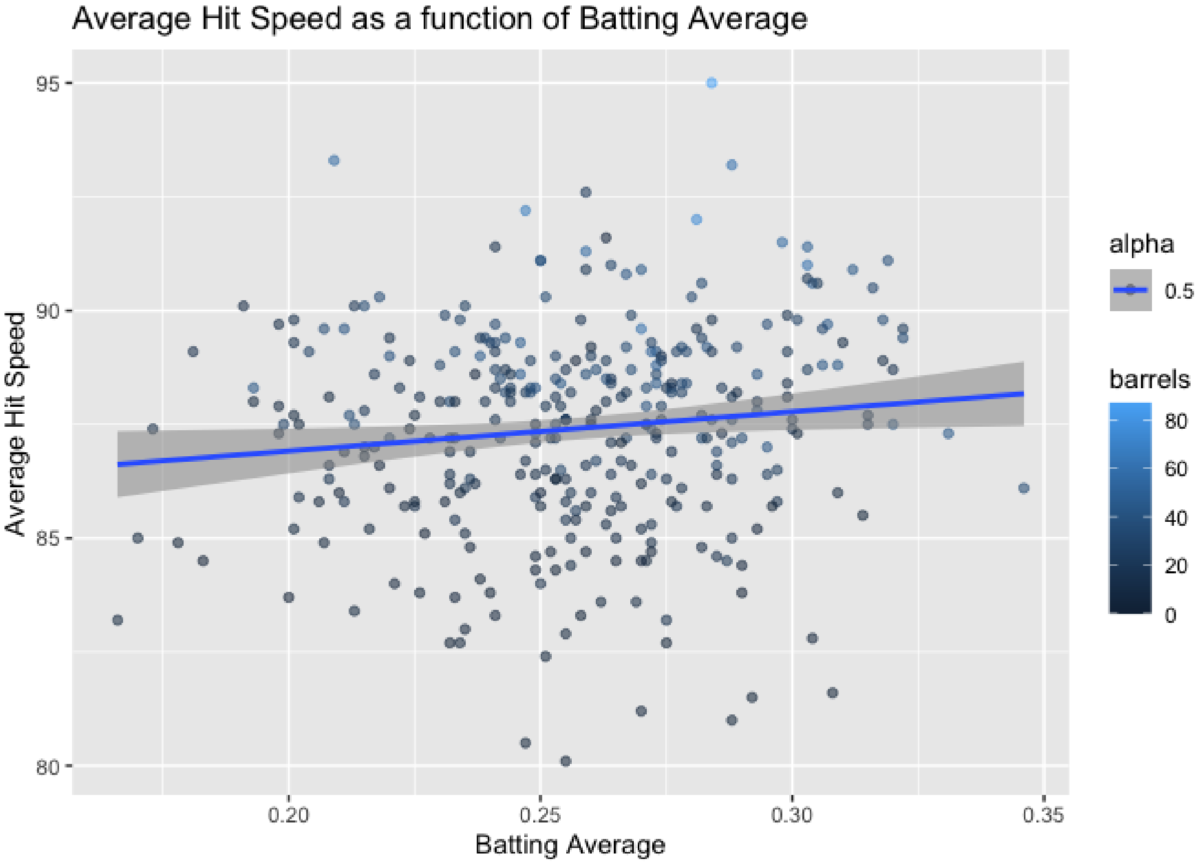
```
glimpse(bsbl)
```

```
## Rows: 907
## Columns: 49
## $ last_name              <chr> "Inciarte", "Lindor", "Strang
## $ first_name             <chr> " Ender", " Francisco", " Dee
## $ player_id              <int> 542255, 596019, 543829, 60514
## $ attempts               <int> 575, 566, 566, 554, 554, 547,
## $ avg_hit_angle          <dbl> 10.2, 14.6, 2.2, 14.3, 8.9, 7
## $ anglesweetspotpercent  <dbl> 34.3, 32.2, 27.7, 31.9, 33.6,
## $ max_hit_speed          <dbl> 102.7, 111.7, 104.1, 111.7, 1
## $ avg_hit_speed          <dbl> 82.8, 89.1, 81.6, 88.4, 88.3,
## $ fbld                   <dbl> 87.5, 92.8, 85.9, 92.9, 90.9,
## $ gb                     <dbl> 78.7, 85.4, 79.6, 87.4, 87.5,
## $ max_distance           <int> 434, 456, 395, 434, 420, 436,
## $ avg_distance           <int> 152, 192, 122, 179, 161, 160,
## $ avg_hr_distance        <int> 382, 401, 380, 396, 393, 403,
## $ ev95plus               <int> 75, 203, 63, 208, 194, 171, 1
## $ ev95percent            <dbl> 13.1, 35.9, 11.2, 37.9, 35.1,
## $ barrels                <int> 3, 40, 1, 25, 21, 23, 13, 10,
## $ brl_percent            <dbl> 0.5, 7.1, 0.2, 4.5, 3.8, 4.2,
## $ brl_pa                 <dbl> 0.4, 5.5, 0.1, 3.5, 3.2, 3.3,
## $ Name                   <chr> "Ender Inciarte", "Francisco 
## $ X                      <int> 3, 2, 8, 4, 26, 12, 41, 14, 1
## $ bbref_id               <int> 572669, 542436, 542583, 57143
## $ season                 <int> 2017, 2017, 2017, 2017, 2017,
## $ Age                    <int> 26, 23, 29, 24, 32, 28, 27, 2
## $ Level                  <chr> "Maj-NL", "Maj-AL", "Maj-NL",
## $ Team                   <chr> "Atlanta", "Cleveland", "Miam
## $ G                      <int> 156, 159, 157, 153, 156, 158,
## $ PA                     <int> 718, 723, 695, 712, 666, 689,
## $ AB                     <int> 662, 651, 653, 628, 620, 643,
## $ R                      <int> 93, 99, 114, 101, 78, 100, 77
## $ H                      <int> 201, 178, 201, 166, 177, 191,
## $ X1B                    <int> 158, 97, 170, 94, 128, 123, 1
## $ X2B                    <int> 27, 44, 20, 46, 30, 44, 38, 2
## $ X3B                    <int> 5, 4, 9, 2, 2, 4, 2, 4, 1, NA
## $ HR                     <int> 11, 33, 2, 24, 17, 20, 14, 8,
## $ RBI                    <int> 57, 89, 33, 102, 85, 88, 69, 
## $ BB                     <int> 49, 60, 25, 77, 36, 38, 47, 5
## $ IBB                    <int> 3, 6, 0, 9, 1, 0, 0, 1, 3, NA
## $ uBB                    <int> 46, 54, 25, 68, 35, 38, 47, 5
## $ SO                     <int> 94, 93, 93, 79, 74, 101, 67, 
## $ HBP                    <int> 0, 4, 10, 2, 2, 3, 3, 6, 1, N
## $ HBP                      <int> 0, 4, 10, 2, 2, 3, 3, 6, 1, N
## $ SH                       <int> 3, 5, 2, 0, 2, 1, 0, 3, 0, NA
## $ SF                       <int> 4, 3, 4, 5, 6, 4, 8, 5, 9, NA
## $ GDP                      <int> 8, 11, 7, 9, 19, 18, 20, 24,
## $ SB                       <int> 22, 15, 60, 26, 1, 25, 19, 6,
## $ CS                       <int> 7, 3, 15, 3, 1, 9, 5, 4, 4, N
## $ BA                       <dbl> 0.304, 0.273, 0.308, 0.264, 0
## $ OBP                      <dbl> 0.350, 0.337, 0.341, 0.344, 0
## $ SLG                      <dbl> 0.409, 0.505, 0.375, 0.459, 0
## $ OPS                      <dbl> 0.759, 0.842, 0.716, 0.803, 0
```

```
# without applying this filter on the data set, there were
# a lot of players with too few plate appearances messing
bsblSub <- bsbl %>% subset(attempts > 100)

bsblSub %>% ggplot(aes(x = anglesweetspotpercent, y = avg_dist
  geom_point() +
  geom_smooth(method = "lm") +
  labs(
    title = "Average Hit Distance as a funciton of launch angl
    x = "Launch Angle Sweet Spot %",
    y = "Average Hit Distance (ft.)"
  )
```



```
# plot batting average by average exit velocity, something tha
# having mergeed the datasets

bsblSub %>% ggplot(aes(x = BA, y = avg_hit_speed, color = barr
  geom_point() +
  geom_smooth(method = "lm") +
  labs (
    title = "Average Hit Speed as a function of Batting Averag
    x = "Batting Average",
    y = "Average Hit Speed"
  )
```



T-Test between the Colorado Rockies Average hit speed and the rest of the MLB

```
# t test
colorado <- bsbl %>% subset(Team == "Colorado")
everyoneElse <- bsbl %>% subset(Team != "Colorado")

t.test(colorado$avg_hit_speed, everyoneElse$avg_hit_speed)
```

```
##
##  Welch Two Sample t-test
##
## data:  colorado$avg_hit_speed and everyoneElse$avg_hit_spee
## t = -2.8622, df = 30.083, p-value = 0.007587
## alternative hypothesis: true difference in means is not equ
## 95 percent confidence interval:
##  -3.3834883 -0.5658663
## sample estimates:
```

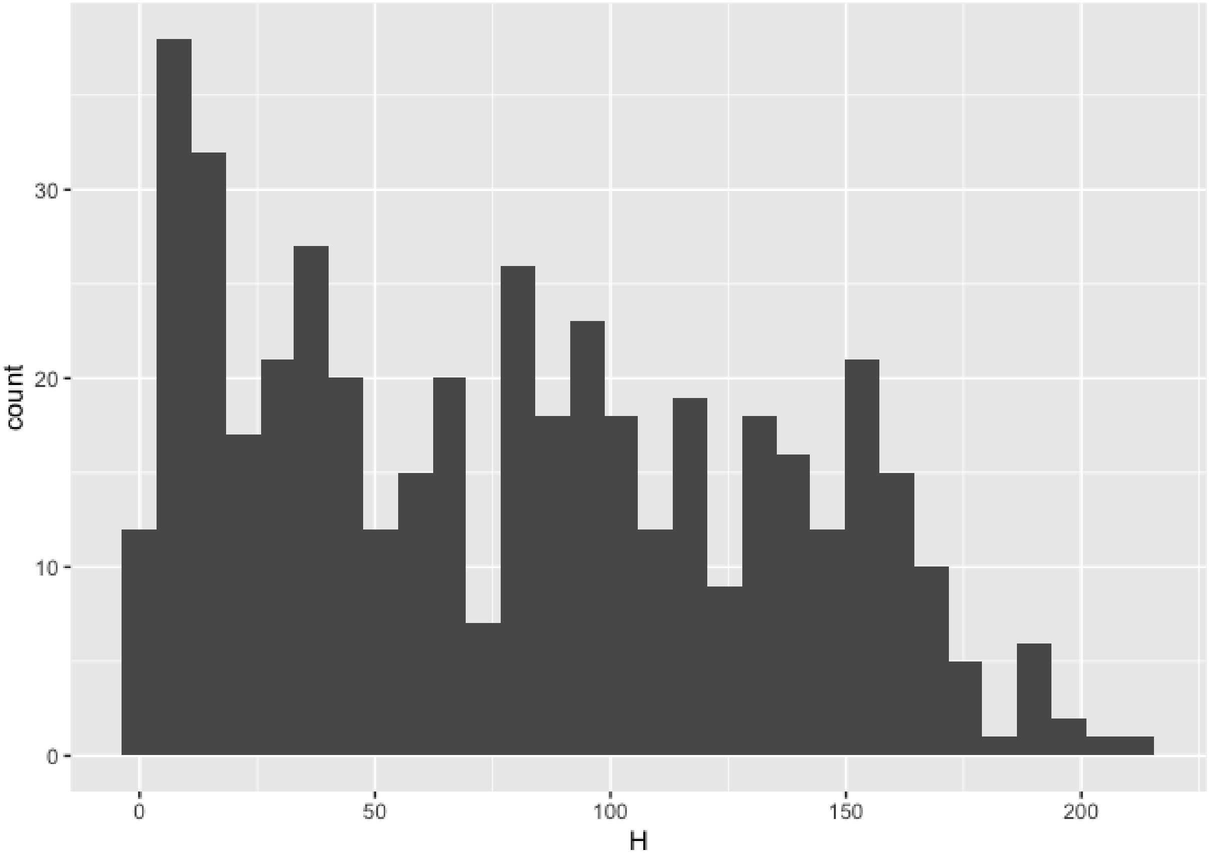
```
## mean of x mean of y
##  83.45600  85.43068
```

Non-Parametric BootStrap

```
bsbl2 <- na.omit(bsbl)

set.seed(10)

x <- c(bsbl2$H)

#define function to calculate mean
meanFunc <- function(x,i){mean(x[i])}

#calculate standard error using 100 bootstrapped samples
boot(x, meanFunc, 100)
```

```
##
## ORDINARY NONPARAMETRIC BOOTSTRAP
##
##
## Call:
## boot(data = x, statistic = meanFunc, R = 100)
##
##
## Bootstrap Statistics :
##     original     bias    std. error
## t1* 79.15639 -0.1551982    2.509104
```

```
bsbl2 %>% ggplot(aes(x = H)) +
 geom_histogram()
```



Parametric BootStrap

```r
# Number of bootstrap samples
B <- 100

# Instantiating matrix for bootstrap samples
paramBoots <- matrix(NA, nrow = length(bsbl2), ncol = B)

# Sampling with replacement B times
for(b in 1:B) {
paramBoots[, b] <- rnorm(n = length(bsbl2), mean = mean(bsbl2$
}

paramBootMedians <- vector(length = B)
for(a in 1:B){
 paramBootMedians[a] <- mean(paramBoots[,a])
}

sd(paramBootMedians)
```

```
## [1] 7.953903
```