Baseball Careers

author

date

Load necessary libraries

```
if(!require(knitr)){install.packages("knitr")}
if(!require(dplyr)){install.packages("dplyr")}
if (!require("readr")) install.packages("readr", dep=FALSE)
if (!require("gridExtra")) install.packages("gridExtra", dep=TRUE)
if (!require("stats")) install.packages("stats", dep=TRUE)
if(!require(FSA)){install.packages("FSA")}
if(!require(ggplot2)){install.packages("ggplot2")}
if (!require("mosaic")) install.packages("mosaic", dep=FALSE)
if (!require("nortest")) install.packages("nortest", dep=TRUE)
if (!require("epitools")) install.packages("epitools", dep=TRUE)
if (!require("prettyR")) install.packages("prettyR", dep=TRUE)
if (!require("rms")) install.packages("rms", dep=TRUE)
# add others as needed
```

Read the baseball data (assuming you have it in a CSV file)

```
baseball <- read.csv("baseball.csv")
dim(baseball)
```

```
## [1] 322  28
```

Data for Problem 1

Comparing American League East vs American League West using 1986 data for seasonal hits

Create a subset of the data for American League

```
problem1 <- baseball %>%
  filter(League == "American") %>%
  select(Salary, Team, Division, YrMajor, logSalary, nAtBat, nBB, nError, nHits, nHome, nRBI, nRuns)

# Frequency table for Division
division_freq <- table(problem1$Division)
print(division_freq)
```

```
##
## East West
##   85   90
```

Sort the data by Division

```
problem1 <- problem1 %>%
  arrange(Division)

# Test for means assuming normal data (t-test)
t_test_result <- t.test(problem1$nHits ~ problem1$Division)
print(t_test_result)
```

```
##
##  Welch Two Sample t-test
##
## data:  problem1$nHits by problem1$Division
## t = 1.6038, df = 164.7, p-value = 0.1107
## alternative hypothesis: true difference in means between group East and group West is not equal to 0
## 95 percent confidence interval:
##  -2.536073 24.483786
## sample estimates:
## mean in group East mean in group West
##           113.3294           102.3556
```

```
# Test for means assuming non-normal data (Wilcoxon test)
wilcoxon_test_result <- wilcox.test(problem1$nHits ~ problem1$Division)
print(wilcoxon_test_result)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  problem1$nHits by problem1$Division
## W = 4288.5, p-value = 0.1669
## alternative hypothesis: true location shift is not equal to 0
```

```
# Test for means assuming non-normal data (KS test)
ks_test_result <- ks.test(problem1$nHits ~ problem1$Division)
print(ks_test_result)
```

```
##
##  Exact two-sample Kolmogorov-Smirnov test
##
## data:  problem1$nHits by problem1$Division
## D = 0.14379, p-value = 0.2608
## alternative hypothesis: two-sided
```

Subset the data for problem 2

Comparing present season number of hits with past average number of hits for teams in National League East

```
prob2 = subset(baseball, Div == "NE")
#summary(prob2)
dim(prob2)
```

```
## [1] 72 28
```

```
#summary(prob2)
x=prob2$nHits                  #present number of hits
y = prob2$CrHits/prob2$YrMajor
```

```
#past average number of hits
```

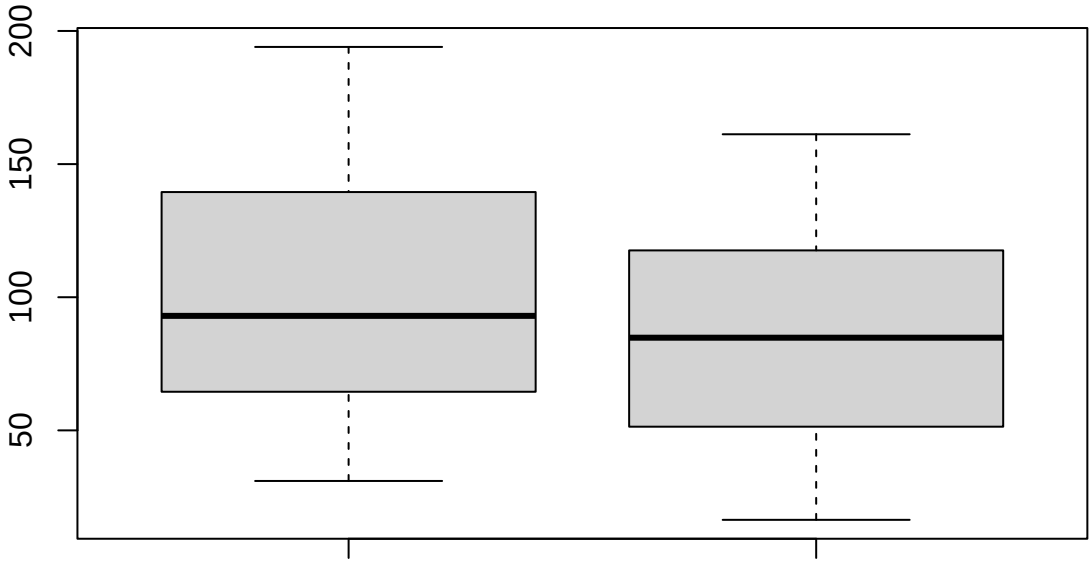
```
favstats(x)
```

```
##  min    Q1 median     Q3 max     mean       sd  n missing
##   31 64.75     93 139.25 194 99.80556 43.80547 72       0
```

```
favstats(y)
```

```
##   min      Q1   median       Q3   max     mean       sd  n missing
##  16.4 51.8125 84.80357 117.5089 161.2 84.38554 39.28305 72       0
```

```
boxplot(x,y, data=prob2)
```




```
t.test(x,y)
```

```
##
##  Welch Two Sample t-test
##
## data:  x and y
## t = 2.2237, df = 140.35, p-value = 0.02776
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   1.710844 29.129191
## sample estimates:
## mean of x mean of y
##  99.80556  84.38554
```

```
wilcox.test(x,y)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  x and y
## W = 3100, p-value = 0.04258
## alternative hypothesis: true location shift is not equal to 0
```

```r
ks.test(x,y)
```

```
##
##  Exact two-sample Kolmogorov-Smirnov test
##
## data:  x and y
## D = 0.19444, p-value = 0.1284
## alternative hypothesis: two-sided
```