MLB Salaries

Thurber

## Get 2017 MLB Salaries

The salaries that are available in ***Salaries2017.CSV*** were downloaded from **USA Today** at <https://www.usatoday.com/sports/mlb/salaries/2017/player/all/>.

 sal2017 = read.csv("Salaries2017.csv", header=TRUE)
 head(sal2017)

## Rank Name Team POS Salary Years Total.Value
## 1 1 Clayton Kershaw LAD SP 33000000 7 (2014-20) 215000000
## 2 2 Zack Greinke ARI SP 31876966 6 (2016-21) 206500000
## 3 3 David Price BOS SP 30000000 7 (2016-22) 217000000
## 4 4 Miguel Cabrera DET 1B 28000000 10 (2014-23) 292000000
## 5 4 Justin Verlander DET SP 28000000 7 (2013-19) 180000000
## 6 6 Jason Heyward CHC RF 26055288 8 (2016-23) 184000000
## Avg.Annual Source
## 1 30714286 https://www.usatoday.com/sports/mlb/salaries/2017/player/all/
## 2 34416666 10/1/2017
## 3 31000000
## 4 29200000
## 5 25714285
## 6 23000000

## Population Values

Numerical and graphical descriptives of the population (2017 MLB players) can be generated. We will focus on the annual salary for the year 2017.

 N = nrow(sal2017)
 summary(sal2017$Salary)

## Min. 1st Qu. Median Mean 3rd Qu. Max.
## 535000 545500 1562500 4468069 6000000 33000000

 mean(sal2017$Salary)

## [1] 4468069

 sd(sal2017$Salary)

## [1] 5948459

 hist(sal2017$Salary)



 boxplot(sal2017$Salary)



## Sample Statistics

Suppose that we now sample from the population as if we were unable to get information for all 868 players who were on the initial 2017 rosters. The sample statistics should be representative of the population values (parameters).

 n = 30
 smplIDs = sample(1:N,n)
 smpl = sal2017[sort(smplIDs),]
 head(smpl)

## Rank Name Team POS Salary Years Total.Value
## 9 9 Jon Lester CHC SP 25000000 6 (2015-20) 1.55e+08
## 11 11 Robinson Cano SEA 2B 24000000 10 (2014-23) 2.40e+08
## 41 41 Ryan Braun MIL OF 19199365 13 (2008-20) 1.50e+08
## 72 72 Giancarlo Stanton MIA RF 14500000 13 (2015-27) 3.25e+08
## 114 114 Madison Bumgarner SF SP 11666667 5 (2013-17) 3.50e+07
## 147 142 Cameron Maybin LAA CF 9000000 1 (2017) 9.00e+06
## Avg.Annual Source
## 9 25833333
## 11 24000000
## 41 11538461
## 72 25000000
## 114 7000000
## 147 9000000

 summary(smpl$Salary)

## Min. 1st Qu. Median Mean 3rd Qu. Max.
## 535000 549250 2312500 5465055 7748050 25000000

 mean(smpl$Salary)

## [1] 5465055

 sd(smpl$Salary)

## [1] 7025281

 hist(smpl$Salary)



 boxplot(smpl$Salary)



 smpl1 = smpl

If we repeat the process, it is unlikely that we will get the same values for the statistics. This is not surprising since it is unlikely that our new sample will contain the same 30 players that the original sample contained.

 n = 30
 smplIDs = sample(1:N,n)
 smpl = sal2017[sort(smplIDs),]
 head(smpl)

## Rank Name Team POS Salary Years Total.Value
## 9 9 Jon Lester CHC SP 25000000 6 (2015-20) 155000000
## 12 11 Prince Fielder TEX DH 24000000 1 (2017) 24000000
## 47 44 Jose Bautista TOR RF 18000000 1 (2017) 18000000
## 49 49 Andre Ethier LAD LF 17500000 5 (2013-17) 85000000
## 109 104 Carlos Santana CLE C 12000000 5 (2012-16) 21001900
## 162 156 Jon Jay CHC CF 8000000 1 (2017) 8000000
## Avg.Annual Source
## 9 25833333
## 12 24000000
## 47 18000000
## 49 17000000
## 109 4200380
## 162 8000000

 summary(smpl$Salary)

## Min. 1st Qu. Median Mean 3rd Qu. Max.
## 535000 547675 2000000 4977655 5487500 25000000

 mean(smpl$Salary)

## [1] 4977655

 sd(smpl$Salary)

## [1] 7059735

 hist(smpl$Salary)



 boxplot(smpl$Salary)



 smpl2 = smpl

As we suspected, the mean of the first sample (5.465055110^{6}) is not equal to the mean of the second sample (4.977655310^{6}). The values are close to each other. They are also close to the population mean (4.468069210^{6}).