Bootstrap Sampling Distributioin

2024-03-29

Stephen Curry is a player in the NBA game data. The file `CurryGameData.csv` contains game information on the 80 games played in the 2021-22 season. - Here we will use the game information for Stephen Curry on the 80 games played in the 2021-22 season. We will consider this to be a population to examine the bootstrap procedure on three attributes; - points per game `pts= (total pts)/(number of games)`, - points and assists per min `ptsAst_min = (ast+pts)/min` and - defensive plays per minutes `defplay_min = (dreb + blk + stl)/min`.

```
games <- read.csv(file = "CurryGameData.csv", header = TRUE)
```

We take the first 40 games as a sample, labelled `gameSample`, was obtained by sampling without replacement.

```
gameSample <- c(1:40)
```

We now write a function named `create.attr3` that takes in a population `pop` and outputs a function which takes in sample indices and outputs the sample estimates for `pts` per game, `ptsAst_min` and `defplay_min`. Apply the population or season `games` to `create.attr3` to create function `attr3` and then apply games 1 to 80 to the function.

```
create.attr3<- function(pop){
  function(sample_indice){
    s = pop[sample_indice,]
    pts = mean(s$pts)
    ptsAst_min = sum(s$ast+s$pts)/sum(s$min)
    defplay_min = sum(s$dreb+s$blk+s$stl)/sum(s$min)

    c(pts, ptsAst_min, defplay_min)
  }
}
attr3 = create.attr3(games)
attr3(1:80)
```

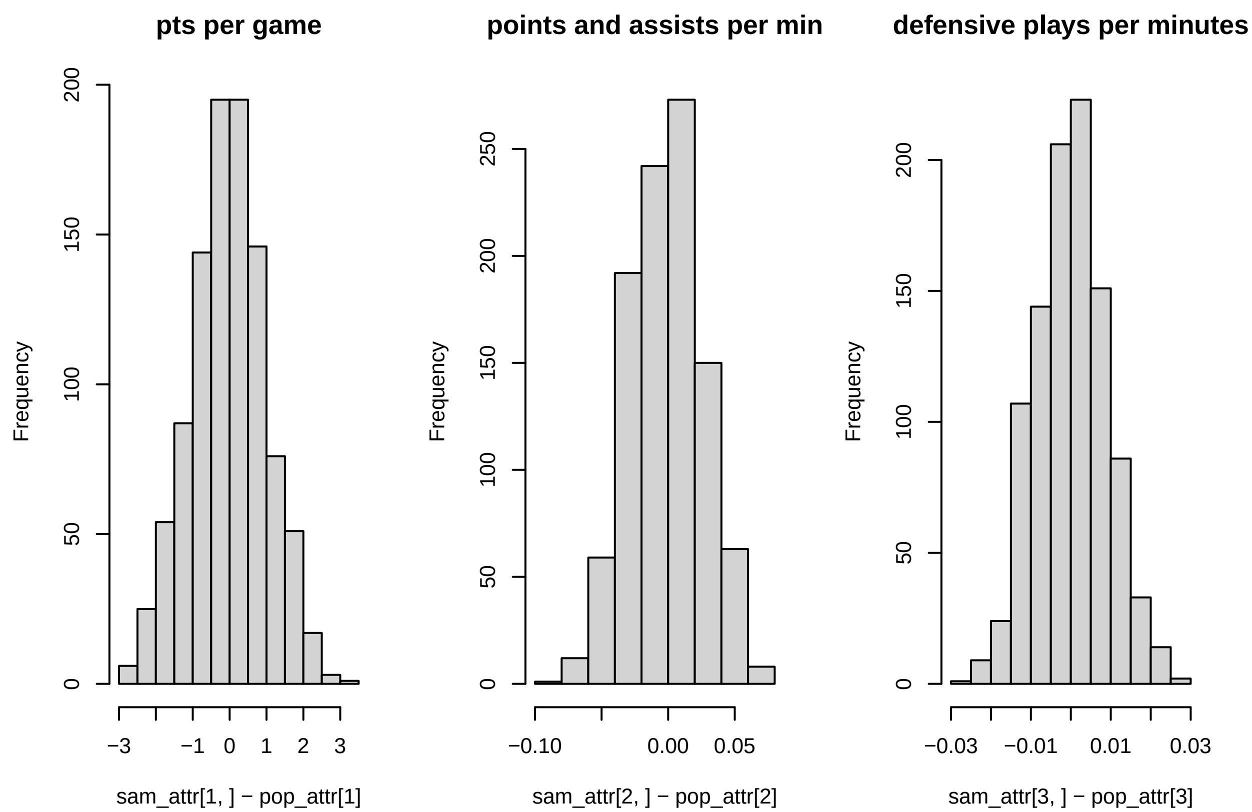
```
## [1] 25.5625000  0.9262087  0.1824791
```

We now look at the Sampling Distribution of the attributes.

We will select $M = 1000$ samples of size $n = 40$ without replacement.Then for each sample, we apply the `attr3` function. Then construct three histograms (in a single row) of the sample error for each attribute.

```
    M=1000
    n=40

    sample<-sapply(1:M, FUN = function(m) sample(1:80, n, replace = FALSE))
```

```r
#sam.coef = replicate(M, {sam = sample(n, 20)})
sam_attr =apply(sample, 2, attr3)

pop_attr = attr3(1:80)

par(mfrow = c(1,3))

hist(sam_attr[1,] - pop_attr[1], main = "pts per game")

hist(sam_attr[2,] - pop_attr[2], main = "points and assists per min")

hist(sam_attr[3,] - pop_attr[3], main = "defensive plays per minutes")
```




````
Now, suppose we do not have access to the entire population of the game sample. We will use the bootstra
We assume we only have the first 40 data points of the population as a sample.

```r
game_sample = c(1:40)
attr3(game_sample)
```

```
## [1] 26.0500000  0.9330909  0.1956364
```
````

By resampling the sample $\mathcal{S}$ with replacement, construct $B = 1000$ bootstrap samples $S_1^\star, S_2^\star, \ldots, S_{1000}^\star$ and calculate the three attributes of interest on each bootstrap sample. Then construct three histograms (in a single row) of the bootstrap sample error for each attribute.

```r
S_star = sapply(1:M, FUN = function(m) sample(1:80, n, replace = FALSE))
bootstrap_attr = apply(S_star, 2, attr3)

par(mfrow = c(1,3))

hist(bootstrap_attr[1,] - attr3(game_sample)[1], main = "pts per game")

hist(bootstrap_attr[2,] - attr3(game_sample)[2], main = "points and assists per min")

hist(bootstrap_attr[3,] - attr3(game_sample)[3], main = "defensive plays per minutes")
```

```
![](Bootstrap-sampling-distribution_files/figure-latex/unnamed-chunk-7-1.pdf)<!-- -->
```

Now, we calculate the standard errors for each sample estimate and then construct a 95% confidence for the population quantity using the percentile method.

```r
standard_errors = apply(bootstrap_attr, 1, sd)


apply(bootstrap_attr, 1, quantile, c(0.025, 0.975))
```

```
##         [,1]      [,2]      [,3]
## 2.5%  23.450 0.8736816 0.1657125
## 97.5% 27.575 0.9781345 0.1989868
```

`The standard errors for pts per game is 1.0443574, the standard errors for points and assists per min`

is 0.027365, the standard errors for defensive plays per minutes is 0.0085337