Homework8

Akhil Havaldar

10/30/2022

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.1.3
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr   0.3.4
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.2      v forcats 0.5.2
```

```
## Warning: package 'ggplot2' was built under R version 4.1.3
```

```
## Warning: package 'tibble' was built under R version 4.1.3
```

```
## Warning: package 'tidyr' was built under R version 4.1.3
```

```
## Warning: package 'readr' was built under R version 4.1.3
```

```
## Warning: package 'dplyr' was built under R version 4.1.3
```

```
## Warning: package 'stringr' was built under R version 4.1.3
```

```
## Warning: package 'forcats' was built under R version 4.1.3
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(readr)
```

Question 1

```r
set.seed(05202001)
delta <- read.csv("Delta delays.csv")
delta <- as.vector(delta$Arrival.Delay)
quants <- quantile(delta)
samp_iqr <- IQR(delta)

B<-10000
boot_samp <- replicate(B, sample(delta, replace=T))
boot_iqr <- apply(boot_samp,2,IQR)

boot_err <- boot_iqr - samp_iqr
boot_err_sort <- sort(boot_err)
p5.0 <- B*0.05
p95.0 <- B*0.95
boot_ci <- samp_iqr - boot_err_sort[c(p95.0,p5.0)]
print(boot_ci)
```

```
## [1] -3.5 21.0
```

- The bootstrap 90% CI for IQR is (-3.5, 21.0)

Question 2

a)

```r
quants <- quantile(delta)
samp_90 <- quantile(delta,c(.9))

B<-10000
boot_samp <- replicate(B, sample(delta, replace=T))
boot_90 <- apply(boot_samp,2,quantile, 0.9)

boot_90 <- boot_90 - samp_90
boot_err_sort <- sort(boot_err)
p025 <- B*0.025
p97.5 <- B*0.975
boot_ci <- samp_iqr - boot_err_sort[c(p97.5,p025)]
print(boot_ci)
```

```
## [1] -8.0 22.5
```

- We get a value of (-8,22.5) for the CI. There is some evidence that the 90th percentile of arrival delay length is a late arrival. ### b)
- No this is not indisputable. For it to be indisputable, the entire confidence interval for the 90th percentile has to be positive.

Question 3

```r
set.seed(05202001)
amer <- read.csv("American delays.csv")
amer <- as.vector(amer$Arrival.Delay)
amer90 <- quantile(amer, c(.90))
print(amer90)
```

```
##  90%
## 33.5
```

```r
print(samp_90)
```

```
##  90%
## 29.4
```

```r
samp_diff <- samp_90 - amer90
print(samp_diff)
```

```
##  90%
## -4.1
```

```r
rand.test <- function(x){
  rand_comb <- sample( c(delta, amer) )
  bquant1 <- quantile(rand_comb[1:x], c(0.90))
  bquant2 <- quantile(rand_comb[(x+1):(length(delta)+length(amer))], c(0.90))
  bquant1 - bquant2
}

boot_diffs_null <- replicate(B, rand.test(length(delta)))
sum(boot_diffs_null <= samp_diff | boot_diffs_null >= 2*mean(boot_diffs_null) - samp_diff)/B
```

```
## [1] 0.8334
```

- Fail to Reject null, which means that there is no evidence that the 90th percentile of arrival delay lengths is different between the two airlines.