Math 152 - Statistical Theory - Homework 6

write your name here

Due: 10/12/2018

Book problems:

8.1 – 1
8.2 – 9, 10
8.4 – 3, 6

- (By pencil/pen/LaTeX) Consider the sample 1,2,3,6 from some distribution.

(a) For one random bootstrap sample, find the probability that the mean is 1.

(b) For one random bootstrap sample, find the probability that the maximum is 6.

(c) For one random bootstrap sample, find the probability that exactly two elements in the sample are less than 2.

R problem (Bootstrapping)

Consider the Flight Data available in the R package `nycflights13` (Airline on-time data for all flights departing NYC in 2013). Consider a subset of the data from only the airlines United and American (note, I've done the data wrangling for you). We will assume the observations represent a random sample from a larger population of UA and AA flights out of NYC. The parameter of interest is the ratio of means of the flight departure delays $\theta = \mu_{UA}/\mu_{AA}$. Consider an estimate of θ to be $\hat{\theta} = \overline{X}_{UA}/\overline{X}_{AA}$.

(a) [I did this for you, but you should recognize the importance of EDA.] Perform some exploratory data analysis [EDA] on the flight delay lengths for each of the UA and AA airlines. Notice the missing values! And the negative numbers. What does it all mean? (Nothing for you to do on this problem, just think about the work below.)

(b) Bootstrap the mean of flight delay lengths for each airline separately, provide plots of the distributions, and describe the distributions in words.

(c) Bootstrap the ratio of the means. Provide plots of the bootstrap distribution and describe the distribution in words.

(d) Recall that the theoretical definition of bias is: $E[\hat{\theta}] - \theta$. Use the bootstrap distribution of the ratio of means to estimate the bias of $\hat{\theta}$. Explain what you see in a sentence or two.

(e) Use the bootstrap distribution to estimate the variability of $\hat{\theta}$. (Use SE with the R function `sd`.) Explain what you see in a sentence or two.

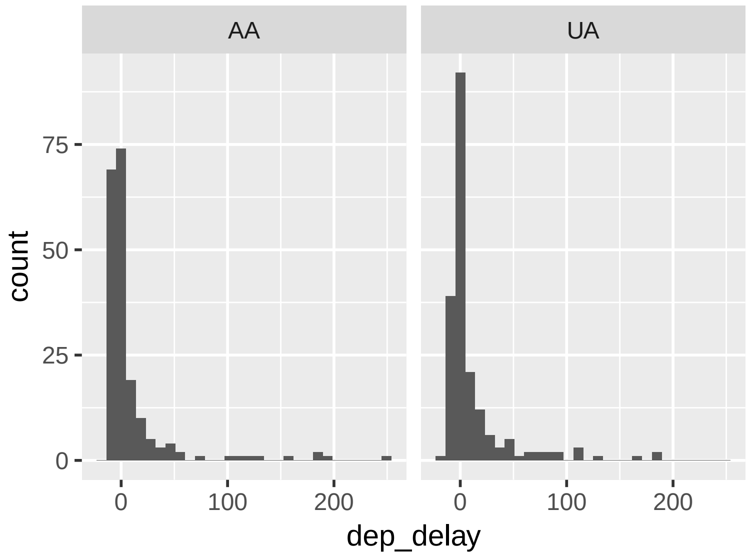
Part (a), no need to add anything here

```r
# Don't need to adjust any of the data wrangling.

library(dplyr)
library(skimr)
library(ggplot2)
library(nycflights13)
data(flights)
```

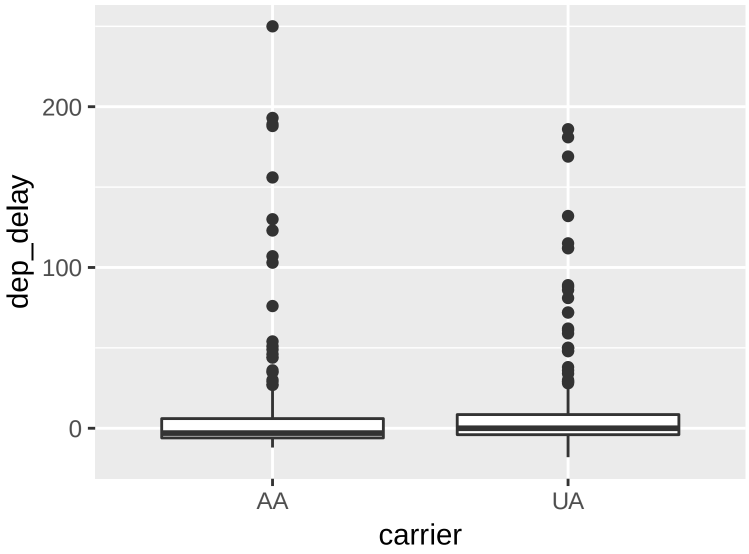
```r
set.seed(4747)
UAAA <- flights %>%
  dplyr::select(dep_delay, carrier) %>%
  dplyr::filter(carrier %in% c("UA", "AA")) %>%
  group_by(carrier) %>%
  sample_n(size = 200) %>%
  ungroup()

# line of code below removes the histogram which makes the file easier to compile
skim_with(numeric = list(hist = NULL))
UAAA %>%
  group_by(carrier) %>%
  skimr::skim(dep_delay)
```

```
## Skim summary statistics
##  n obs: 400
##  n variables: 2
##  group variables: carrier
##
## -- Variable type:numeric ------------------------------------------------------
##  carrier  variable missing complete   n  mean    sd  p0 p25 p50 p75 p100
##       AA dep_delay       4      196 200  8.96 37.2  -12  -6  -3 6    250
##       UA dep_delay       5      195 200 11.21 32.69 -18  -4   0 8.5  186
```

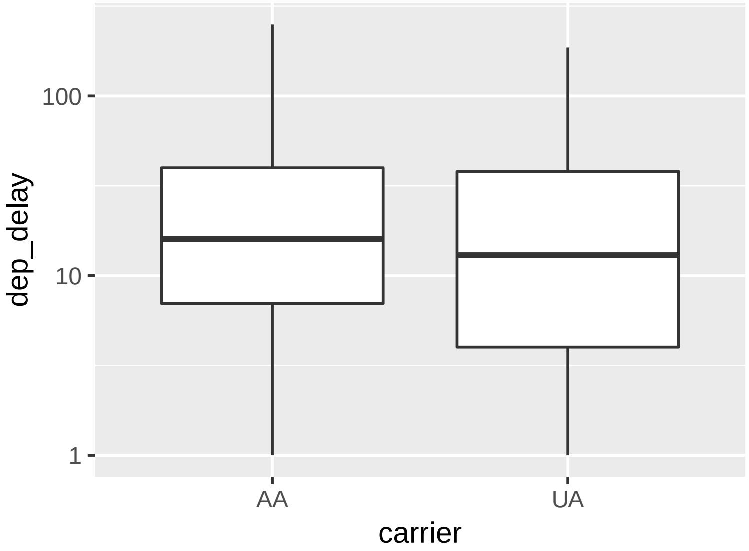
```r
# Removing the missing values
UAAA <- UAAA %>% dplyr::filter(!is.na(dep_delay))
UA_delay <- UAAA %>% dplyr::filter(carrier == "UA") %>% dplyr::select(dep_delay) %>% pull()
AA_delay <- UAAA %>% dplyr::filter(carrier == "AA") %>% dplyr::select(dep_delay) %>% pull()

ggplot(UAAA, aes(x = dep_delay)) + geom_histogram(bins=30) + facet_grid(~carrier)
```



```r
ggplot(UAAA, aes(x = carrier, y = dep_delay)) + geom_boxplot()
```



```
UAAA %>%
  dplyr::filter(dep_delay > 0) %>%
  ggplot(aes(x = carrier, y = dep_delay)) + geom_boxplot() + scale_y_log10()
```



Part (b)

```
# reps is the number of bootstraps, "B"
reps <- 1000
UA_bs <- numeric(reps)
AA_bs <- numeric(reps)

for (i in 1:reps){
  UA_bs[i] <- mean(sample(UA_delay, replace = TRUE))
  AA_bs[i] <- mean(sample(AA_delay, replace = TRUE))
}

# make one histogram for each set of bootstrapped statistics (two histograms total)
```

Parts (c), (d), (e)

```
# reps is the number of bootstraps, "B"
reps <- 1000

# place holder for ONE stat of interest


for (i in 1:reps){

}

# one histogram describing the sampling distribution of the ratio statistic


# estimate the bias


# estimate the variability
```