Math 152 - Statistical Theory - Homework 6

write your name here

Due: not due ever

Important Note:

You should work to turn in assignments that are clear, communicative, and concise. Part of what you need to do is not print pages and pages of output. Additionally, you should remove these exact sentences and the information about HW scoring below.

Click on the *Knit to PDF* icon at the top of R Studio to run the R code and create a PDF document simultaneously. [PDF will only work if either (1) you are using R on the network, or (2) you have LaTeX installed on your computer. Lightweight LaTeX installation here: https://yihui.name/tinytex/]

> Either use the college's RStudio server (https://rstudio.pomona.edu/) or install R and R Studio on to your personal computer. See: https://research.pomona.edu/johardin/math152f20/ for resources.

Assignment

1: PodQ

Describe one thing you learned from someone in your pod this week (it could be: content, logistical help, background material, R information, etc.) 1-3 sentences.

2: 8.1.1

Suppose that a random sample $X_1, \ldots, X_n$ is to be taken from the uniform distribution on the interval $[0, \theta]$ and that θ is unknown. How large a random sample must be taken in order that

$$P(|\max\{X_1, \ldots, X_n\} - \theta| \leq 0.1\theta) \geq 0.95,$$

for all possible θ?

3: 8.1.6

For the conditions of Exercise 5, use the central limit theorem in Sec. 6.3 to find approximately the size of a random sample that must be taken in order that

$$P(|\overline{X} - p| \leq 0.1) \geq 0.95 \text{ when } p = 0.2.$$

4: 8.2.9

Suppose that $X_1, \ldots, X_n$ form a random sample from the normal distribution with mean μ and variance σ^2. Find the distribution of

$$\frac{n(\overline{X} - \mu)^2}{\sigma^2}.$$

5: 8.2.10

Suppose that six random variables $X_1, \ldots, X_6$ form a random sample from the standard normal distribution, and let

$$Y = (X_1 + X_2 + X_3)^2 + (X_4 + X_5 + X_6)^2.$$

Determine a value of c such that the random variable cY will have a χ^2 distribution.

6: 8.4.3

Suppose that the five random variables $X_1, \ldots, X_5$ are i.i.d. and that each has the standard normal distribution. Determine a constant c such that the random variable

$$\frac{c(X_1 + X_2)}{(X_3^2 + X_4^2 + X_5^2)^{1/2}}$$

will have a t distribution.

7: 8.4.6

In Example 8.2.3, suppose that we will observe n = 20 cheese chunks with lactic acid concentrations $X_1, \ldots, X_{20}$. Find a number c so that

$$P(X_{20} \leq \mu + c\sigma') + 0.95.$$

Note that σ' is the estimator such that $(\sigma')^2$ is the unbiased estimator of of σ^2 as defined on page 482, equation 8.4.5.

8: by pencil (LaTeX)

Consider the sample (observed data) 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: bootstrapping the flights data

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.

Some R code you might find useful:

a. No need to add or change anything here

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

library(tidyverse)
library(skimr)
library(nycflights13)
data(flights)
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()


UAAA %>%
  group_by(carrier) %>%
  skimr::skim_without_charts(dep_delay)
```

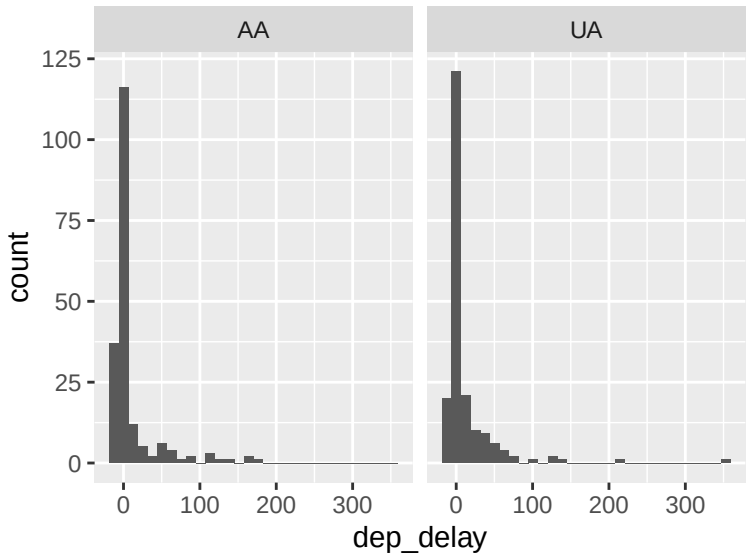
Table 1: Data summary

Name	Piped data
Number of rows	400
Number of columns	2
_____	
Column type frequency:	
numeric	1
_____	
Group variables	carrier

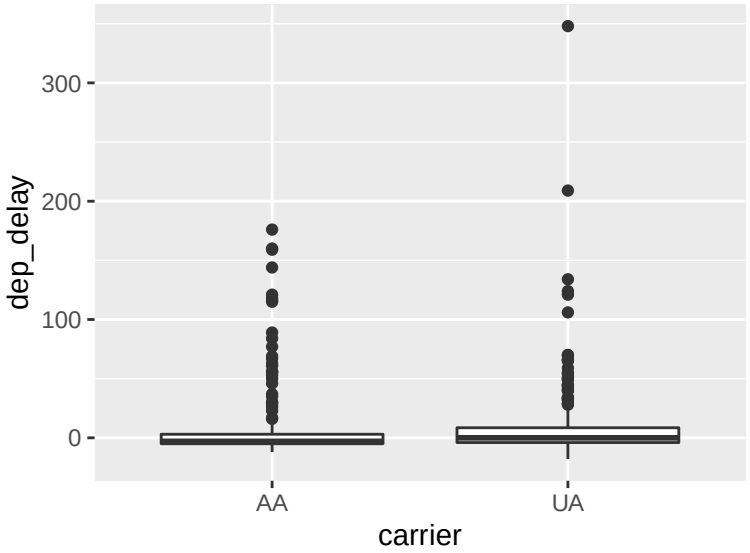
Variable type: numeric

skim_variable	carrier	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100
dep_delay	AA	7	0.96	8.82	33.07	-12	-5	-3	3.0	176
dep_delay	UA	1	1.00	11.11	36.60	-18	-4	0	8.5	348

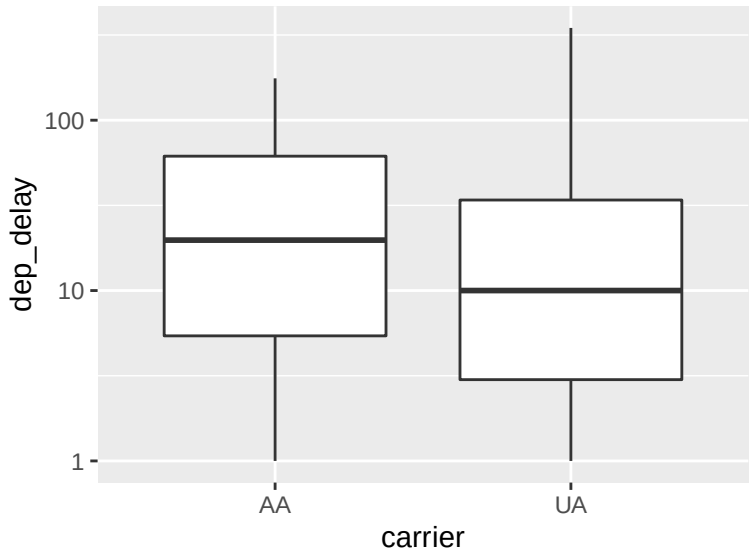
```
# Removing the missing values
UAAA <- UAAA %>% dplyr::filter(!is.na(dep_delay))
UA_delay <- UAAA %>% dplyr::filter(carrier == "UA") %>% dplyr::select(dep_delay) %>% pull()
AA_delay <- UAAA %>% dplyr::filter(carrier == "AA") %>% dplyr::select(dep_delay) %>% pull()

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




```
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()
```



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)
```

Use the code below for c., d., and 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
```