514 Lab 3

Julio Pagan, Joseph Fulkerson

Due Date 10/13

Packages

```
# add packages you need for this assignment
library(tidyverse)
library(tigerstats)
```

How to work with this document

There are several ways to format your answer, but whatever you do, please make sure it is *readable* by a human and clean. That is, don't leave stray comments and commented instructions in your submitted work. You may use a quote code > to start your answer after any r/python code chunks you are typing. An example is provided in the first question below.

Part ONE: Multiple Choice and TRUE/FALSE (15 points)

Question 1: (3 points) To find a confidence interval on population mean *when population variance is known*, which of the following should we use?

(In this part, suppose X_1, . . . , X_1000 is a random sample (of size 1000) from some **unknown** distribution.)

A. The normal distribution (with the Z statistic)

B. The normal distribution (with the Z statistic), but ONLY if X comes from a normal distribution

C. The t-distribution (with the T statistic)

D. The t-distribution (with the T statistic), but ONLY if X comes from a normal distribution

> Group13 answer: A

Question 2: (3 points) To find a confidence interval on population mean *when population variance is unknown*, which of the following should we use?

(In this part, suppose X_1, . . . , X_1000 is a random sample (of size 1000) from some unknown distribution.)

A. The normal distribution (with the Z statistic)

B. The normal distribution (with the Z statistic), but ONLY if X comes from a normal distribution

C. The t-distribution (with the T statistic)

D. The t-distribution (with the T statistic), but ONLY if X comes from a normal distribution

Group13 answer: D

Question 3: A summary of one numerical variable is as follows. Which of the following are TRUE/FALSE? Explain it. (9 points, 3 points each question)

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  4.300   5.100   5.800   5.843   6.400   7.900
```

A. 50% values of this variable are greater than 5.8433333.

Group13 answer: FALSE (50% of the values are greater than the `median` NOT the `mean`)

B. The middle 50% values of this variable between approximately 5.1 and 6.4.

Group13 answer: TRUE (%50 of the middle variable is greater than the median of 5.1 and 6.4)

C. The smallest value of this varaible is 4.3 and the largest value is 7.9.

Group13 answer: TRUE (`min` returns the smallest value and `max` returns the largest value)

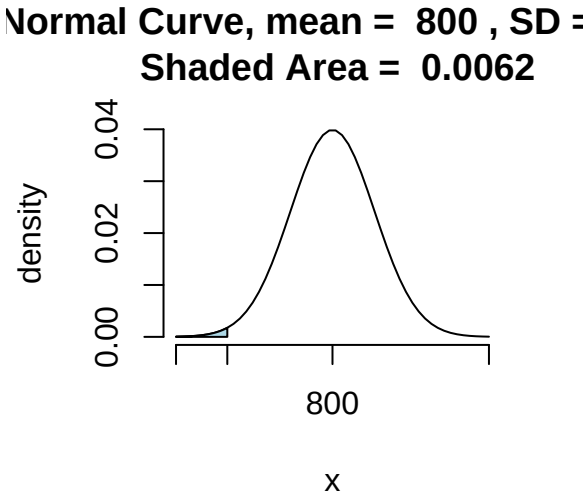
Part Two: Construct Confidence Interval (30 points)

Pro-tip: You may use either R or Python or hand-calculations to answer the computational part of this question, however, you do need to–in any case–explain and justify your answer.

Problem 1. An electrical firm manufactures light bulbs that have a length of life that is approximately normally distributed, with mean equal to 800 hours and a standard deviation of 40 hours. A random sample of 16 bulbs will have an average life of less than 775 hours. (15 points)

```
pnormGC(775, mean = 800, sd = 40/sqrt(16), graph = TRUE)
```

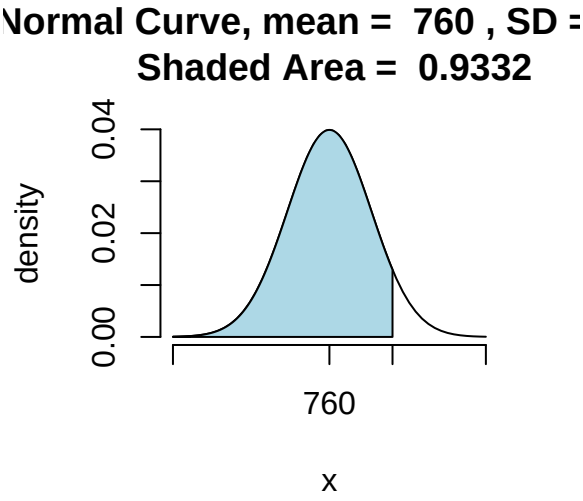
a. Give a probabilistic result that indicates how rare an event $\bar{X} \leq 775$ is when $\mu = 800$. (Hint: Calculate the probability $P(\bar{X} \leq 775)$ when $\mu = 800$). On the other hand, how rare would it be



if μ truly were, say, 760 hours?

```
[1] 0.006209665
```

```
pnormGC(775, mean = 760, sd = 40/sqrt(16), graph = TRUE)
```




```
[1] 0.9331928
```

Group13 answer: By Central Limit Theory, we know when n is large. Based on the sample, it is more likely that μ would be 760 instead of 800

```r
interval <- 0.05
qnormed <- qnorm(1 - interval / 2)
sd <- 40
n <- 16
x_bar <- 775
lo_bd <- x_bar - qnormed * sd / sqrt(n)
lo_bd
```

b. Please construct a 95% confidence interval on μ with $\bar{X} = 775$. Is 800 inside the interval?

```
[1] 755.4004
```

```r
up_bd <- x_bar + qnormed * sd / sqrt(n)
up_bd
```

```
[1] 794.5996
```

Group13 answer: 800 is not inside an interval of 95% confidence

Problem 2. A maker of a certain brand of low-fat cereal bars claims that the average saturated fat content is 0.5 gram. In a random sample of 8 cereal bars of this brand, the saturated fat content was 0.6, 0.7, 0.7, 0.3, 0.4, 0.5, 0.4, and 0.2. *Assume a normal distribution.* (15 points)

```r
sat_fat <- c(0.6, 0.7, 0.7, 0.3, 0.4, 0.5, 0.4, 0.2)
t.test(sat_fat, alternative = c("two.sided"), mu = 0.5, conf.level = 0.95)
```

a. Please construct a 95% confidence interval on the average saturated fat content.

```
    One Sample t-test

data:  sat_fat
t = -0.38592, df = 7, p-value = 0.711
alternative hypothesis: true mean is not equal to 0.5
95 percent confidence interval:
 0.32182 0.62818
sample estimates:
mean of x
    0.475
```

b. Would you agree with the claim? Justify your answer.

Group13 answer: Usint a t-Test, we conclude that we should not reject the claim because the p value is much larger than 0.05 and the 95% confidence interval is (0.32182, 0.6818) with the sample average being 0.475

Part Three: More EDA Practice (50 points)

Instructions: Please review EDA Handout first. Import the needed packages first.

- Obtaining the adult dataset

Tasks

For the following exercises, work with the `adult.data` data set. Use either `Python` or `R` to solve each problem. Please read the `adult.name` file to understand each attribute.

```
setwd(getwd())
adult <- read.csv("adult.data", sep = ",")
```

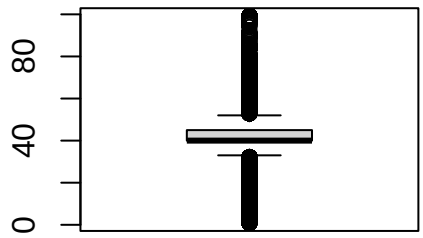
a. Import the `adult.data` data set and name it `adult`. (5 points)

```
names(adult)[1:15] <- c("age","workclass", "fnlwgt",
  "education","education-num","marital-status","occupation",
  "relationship","race","sex","capital-gain","capital-loss","hours-per-week",
  "native-country","class(response)")
hrs_per_wk <- scale(adult$`hours-per-week`)
outliers <- adult[hrs_per_wk < -3 | hrs_per_wk > 3,]
#too many outliers, just show the first six records
head(outliers)
```

b. Standardize `hours-per-week` and indicate if there is any outlier (5 points)

```
    age         workclass fnlwgt     education education-num
10   37           Private 280464  Some-college            10
28   39           Private 367260       HS-grad             9
77   67                 ? 212759          10th             6
157  71  Self-emp-not-inc 494223  Some-college            10
189  58         State-gov 109567     Doctorate            16
272  50  Self-emp-not-inc  30653       Masters            14
         marital-status       occupation   relationship   race   sex
10   Married-civ-spouse  Exec-managerial        Husband  Black  Male
28            Divorced  Exec-managerial  Not-in-family  White  Male
77   Married-civ-spouse                ?        Husband  White  Male
157           Separated            Sales      Unmarried  Black  Male
189  Married-civ-spouse   Prof-specialty        Husband  White  Male
272  Married-civ-spouse  Farming-fishing        Husband  White  Male
    capital-gain capital-loss hours-per-week native-country class(response)
10             0            0             80  United-States            >50K
28             0            0             80  United-States           <=50K
77             0            0              2  United-States           <=50K
157            0         1816              2  United-States           <=50K
189            0            0              1  United-States            >50K
272         2407            0             98  United-States           <=50K
```

```
boxplot(adult$`hours-per-week`)
```



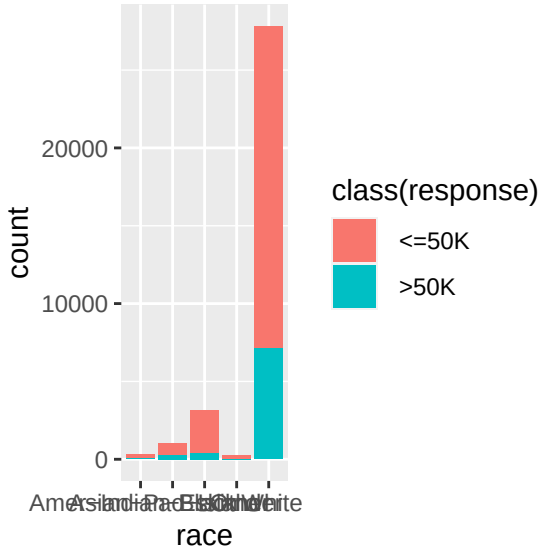
Group13 answer: There are a a lot of outliers based on the hours-per-week

```
table(adult$race, adult$`class(response)`)
```

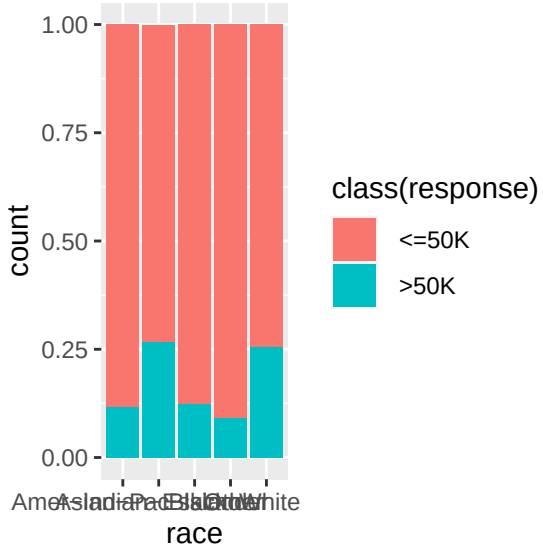
c. Show a bar graph of `race` with a response `class` overlay. What conclusion can you draw from the bar graph? (10 points)

```
                     <=50K  >50K
   Amer-Indian-Eskimo   275    36
   Asian-Pac-Islander   763   276
   Black               2737   387
   Other                246    25
   White              20698  7117
```

```
ggplot(data = adult, mapping = aes(x = race, fill = `class(response)`)) +
  geom_bar()
```



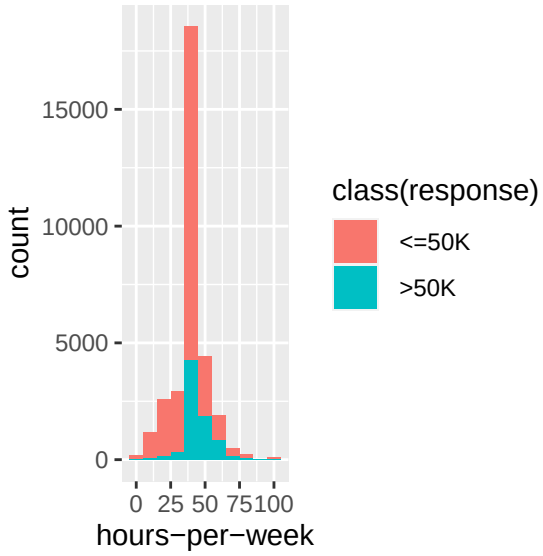

```
ggplot(data = adult, mapping = aes(x = race, fill = `class(response)`)) +
  geom_bar(position = "fill")
```



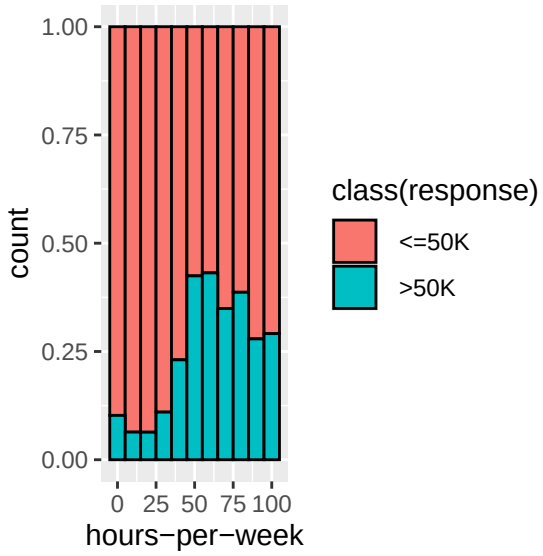
Group13 answer: Whithout `fill`, `white` race attribute dominates. Using `fill` we can see that the percentage of `>50K` in `Asian-Pac-Islander` is closer to the distribution of the `white` race. The `other` category has the most scewed ratio with the least % of `>50K`

```
ggplot(data = adult,
  mapping = aes(x = `hours-per-week`, fill = `class(response)`)) +
  geom_histogram(binwidth = 10)
```

d. Select any numeric attribute and show a histogram of it with a response `class` overlay. What conclusion can you draw from the histogram? (10 points)



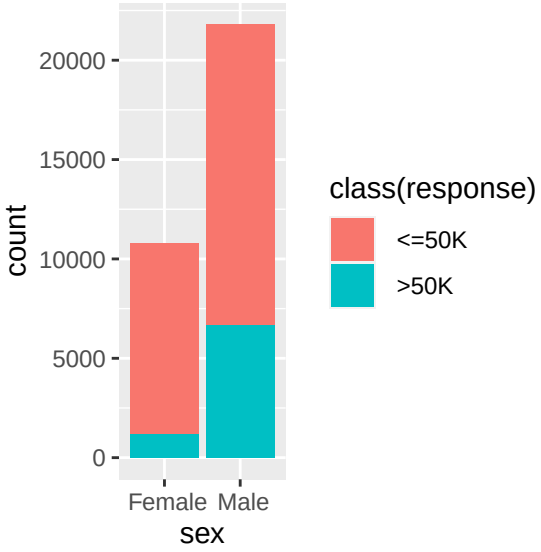

```
ggplot(data = adult,
  mapping = aes(x = `hours-per-week`, fill = `class(response)`)) +
  geom_histogram(binwidth = 10, color = "black", position = "fill")
```



Group 13 answer: Histogram of `hours-per-week` shows that the mayority of adults work aprox. 40 hours per week. Using `fill` we can see the ratio of adults earning `<=50` or `>50K` which shows that most of the adults making `>50K` work over 40 hours per week. Only a minority are able to earn `>50k` by working less than 35 hours per week.

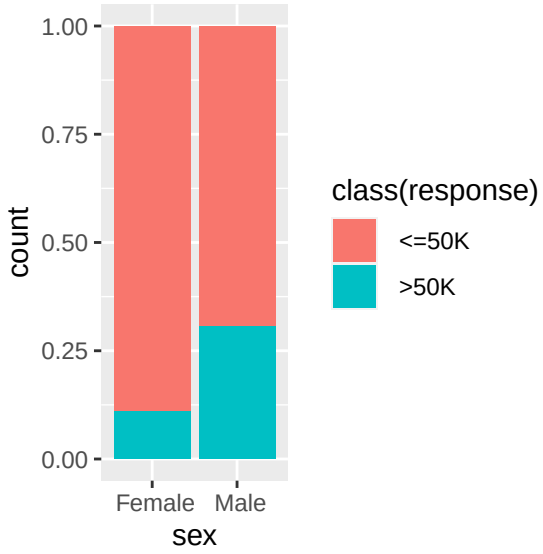
```
ggplot(data = adult, mapping = aes(x = `sex`, fill = `class(response)`)) +
  geom_bar()
```

e. Select any two attributes and show a plot, what conclusion can you draw from the plot?



(10 points)

```
ggplot(data = adult, mapping = aes(x = `sex`, fill = `class(response)`)) +
  geom_bar(position = "fill")
```

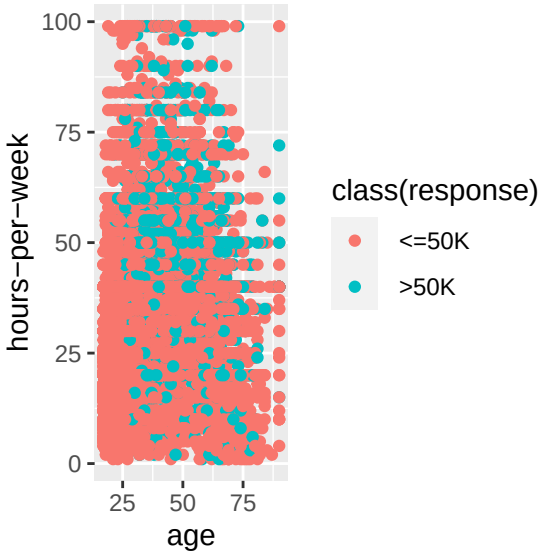


Group13 answer: After analyzing a plot of `class` and `sex` we can conclude that there are significally more adult men in the class of `>50k` in comparisson to females.

```
ggplot(data = adult) +
  geom_point(mapping = aes(x = age, y = `hours-per-week`,
  colour = `class(response)`))
```

f. Select any three attributes and plot their relationship using 2D scatter plot, use one of the selected attributes as the color code when plotting, what can you say about the cor-

relation of these attributes? What conclusion can you draw from the plot? (10 points)



Group13 answer: Using `age` and `hours-per-week` as the main attributes in the scatterplot, color coded according to their `class` attribute, we conclude that most people in the class of `>50k` are within an approximate age range of 33 and 65