HW3_Vikas Sanil

Vikas Sanil

Due Date 2/28

Packages

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

Part ONE: Review the approach to location and scale problems for one and two populations (10 points, 5 points each question)

- 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

  Answer: A. The normal distribution (with the Z statistic)

- 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

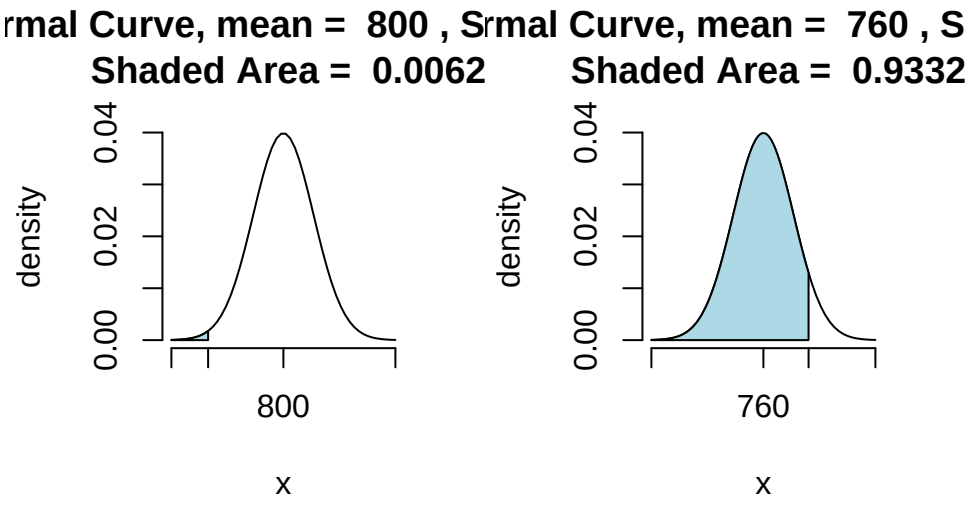
  Answer: D. The t-distribution (with the T statistic), but ONLY if X comes from a normal distribution

Part Two: Confidence Interval (20 points)

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. (10 points)

a. Give a probabilistic result that indicates how rare an event $\bar{X} \leq 775$ is when $\mu = 800$. On the other hand, how rare would it be if μ truly were, say, 760 hours?

```
nsize<-16
xbar<-775
mu1<-800
std<-40
mu2<-760
standardDev<-std/sqrt(nsize)
par(mfrow=c(1,2))
p1<-pnormGC(xbar, region="below", mean = mu1, sd= standardDev, graph = TRUE )
p2<-pnormGC(xbar, region="below", mean = mu2, sd= standardDev, graph = TRUE )
```



Answer:

- The probability of event $\bar{X} \leq 775$ when $\mu = 800$ is 0.0062097.
- The probability of event $\bar{X} \leq 775$ when $\mu = 760$ is 0.9331928.

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

```
xbar<-775
std<-40
nsize<-16
alpha<-.05
loBound<-xbar - qnorm(1-alpha/2)*(std/sqrt(nsize))
loBound
```

```
[1] 755.4004
```

```
upBound<-xbar + qnorm(1-alpha/2)*(std/sqrt(nsize))
upBound
```

```
[1] 794.5996
```

Answer:

The 95% confidence interval on μ with $\bar{X} = 775$ falls between 755.4003602 and 794.5996398. Hence 800 is outside the 95% confidence interval.

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.* (10 points)

```
sampleCerealBars<-c(0.6, 0.7, 0.7, 0.3, 0.4, 0.5, 0.4,0.2)
t.test(sampleCerealBars, conf.level = .95)
```

```
    One Sample t-test

data:  sampleCerealBars
t = 7.3325, df = 7, p-value = 0.0001583
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 0.32182 0.62818
sample estimates:
mean of x
    0.475
```

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

Answer: The 95% confidence interval on the average satuarted fat content is between 0.32182 gram and 0.62818 gram.

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

Answer: The average stated is well within the confidence interval of the random sample tested and closer to the random sample average.But the sample size is too low. More random sample is required to confirm the claim.

Part Three: Working With Data (65 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.

a. Import the `adult.data` data set and name it `adult`. (10 points)

```
url <- "http://archive.ics.uci.edu/ml/machine-learning-databases/adult/adult.data"
adult <- read.csv(url, strip.white = TRUE, header = FALSE)
renameCol<-c("age", "Workclass", "fnlwgt", "education", "educationnum", "maritalstatus", "occupation",
colnames(adult)<-renameCol
```

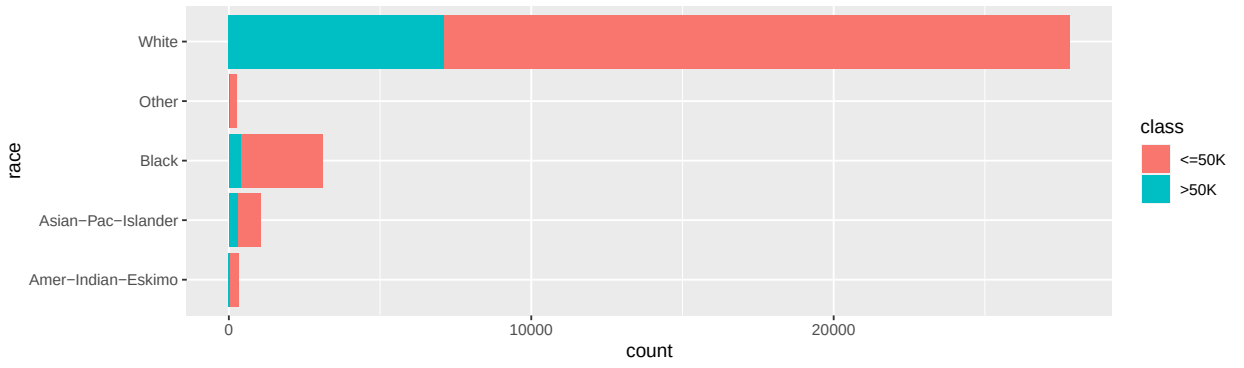
b. Standardize `hours-per-week` and indicate if there is any outlier (10 points)

```
hoursPerWeek_z<-scale(adult$hoursperweek)
hoursPerWeek_outliers<-adult[hoursPerWeek_z<-3|hoursPerWeek_z>3,]
numHoursPerweekOutlier<-nrow(hoursPerWeek_outliers)
```

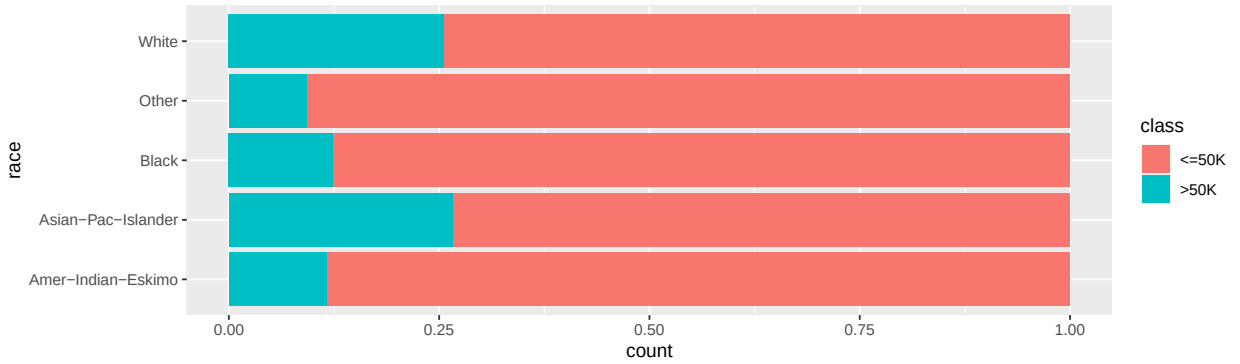
Answer: The number of outlier in adult dataset when 'hours-per-week' standardize is 32561.

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

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




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

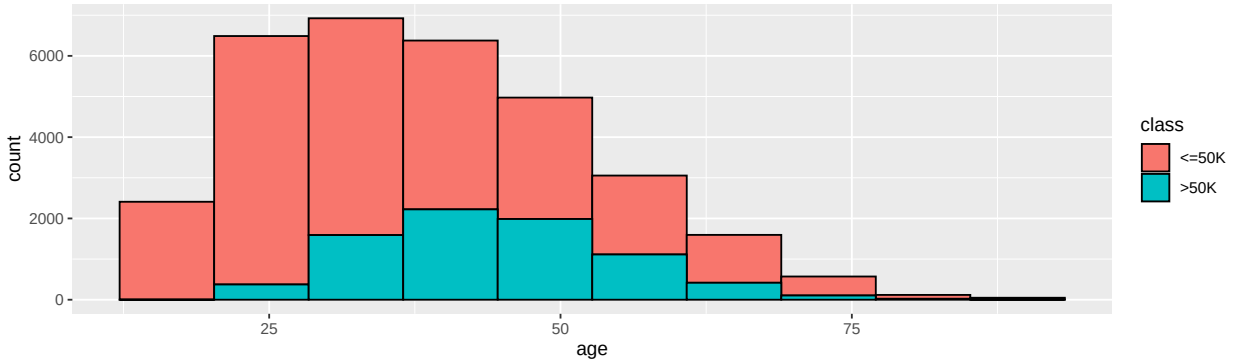


Answer: Based on above plots we can conclude:

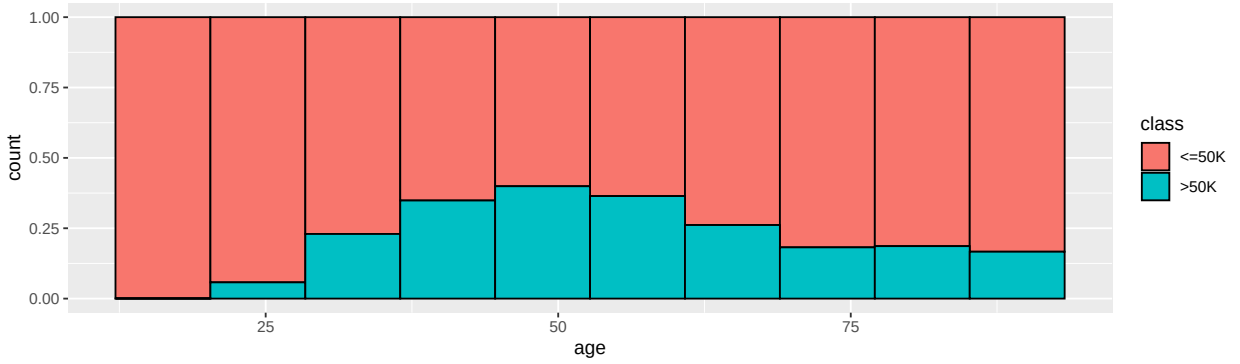
- 'White' race is majority in both class - '<=50K' and '>50K' in the dataset.
- According to dataset, 'Asian-Pac-Islander' race has higher probability of gettign into '>50K' class as 'White' race.

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, aes(x = age))+
  geom_histogram(aes(fill= class), bins=10, color ="black")
```




```
ggplot(data=adult, aes(x = age))+
  geom_histogram(aes(fill= class), bins=10, color ="black", position="fill")
```



Answer: Based on above plots we can conclude:

- Large number of adults are in <=50K class.
- Around age 50 the probability of being in '>50k' class is high.

e. Select any two attributes and show a plot, what conclusion can you draw from the plot? (10 points)

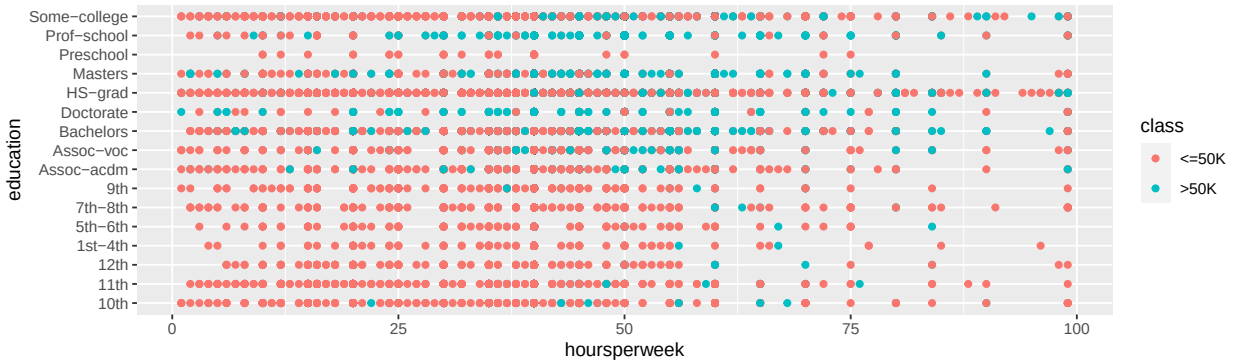
```
ggplot(data=adult)+
  geom_bin2d(mapping = aes( x= Workclass, y = age))
```

Answer: Based on above plot we can conclude:

- More Adult join 'Private' work-class.
- Never-worked work-class has only younger adults.
- The Work-class related to Goverment(Local/State/Fedral) doesn't has older adult, which shows strict retirement policy.

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 correlation of these attributes? What conclusion can you draw from the plot? (15 points)

```
ggplot(data=adult, aes(x= hoursperweek, y= education  , colour=class))+
  geom_point()
```



Answer: Based on above plot we can conclude:

- Lower education level has very few >50k class.
- The number of '>50K' class is relatively higher for education levels like -Doctorate, Masters, Bachelor and Prof-School.