Problem Set 1

Kartik Srikumar

Collaborators: N/A

Some Useful Reminders.

- Be sure to frequently knit your work, such as after finishing each problem. This makes it easier to diagnose the source of any knitting errors.
- Check the Common R and R Markdown Errors document linked to under Module 0 for help troubleshooting error messages.
- For additional help with knitting errors, post on the #r_help Slack channel or bring questions to office hours.

Problem 1.

a) The scenario states that the sample studied consisted of patients with poor heart health and scheduled to undergo heart examinations. There is a sampling issue here in that the sample does not properly represent the population. If only people with poor heart health are studied, it is very likely that the rate of heart attacks will be high in general for those people. A better way to sample the population would be to use *Stratified Sampling*: divide the population into strata such as people with no heart issues, people with heart issues, people interested in the stock market, etc. and then draw a random sample from each stratum.

b) In the scenario, the sample was drawn from a group of homeless people who received medical attention from a clinic. This sample is not representative of the population (homeless people in this case), because a *Convenience Sample* has been used since they are easily accessible. The argument provided by the authors is not as relevant, because access to the clinic does not imply willingness, intent or action on the part of the homeless people experienceing mental health issues. A better way to sample would be to draw a random sample from all homeless people.

Problem 2.

a) One key takeaway from the description of the study is that there may be other things apart from coffee consumption that are linked to the risk of colorectal cancer such as fruit and vegetable consumption. This means that the folks who reported drinking coffee could have also been consuming high amounts of fruits and vegetables. The consumption of fruits and vegetables becomes a confounding variable in this case, which means that it could be the reason for higher survival rather than the coffee consumption itself. The author is correct because all that the study result can claim is that "Coffee does not seem to increase the risk of death in patients with colorectal cancer." Additionally, since the study only consisted of patients with advanced stage of colorectal cancer, no conclusions can be drawn about preventing the cancer in the first place.

b) One way of understanding the effect of coffee consumption on developing colorectal cancer is to use an experiment approach instead of an observational study, since observational studies are not the best to make causal conclusions between explanatory and response variables. For the experiment, we could use a control group which consists of people without colorectal cancer, who are not allowed to drink coffee and an experimental group of the same type of people, but who are asked to consume a certain amount of coffee regularly. Over the course of time, the number of people diagnosed with colorectal cancer can be observed and conclusions may be drawn. One important note is to be able to ensure that the people in both groups have a similar mix of dietary habits (fruit and vegetable consumption etc.).

c) One major disadvantage is that the above experiment is expensive in that it is costly, involved and time-consuming. It is much easier to conduct an observational survey.

Problem 3.

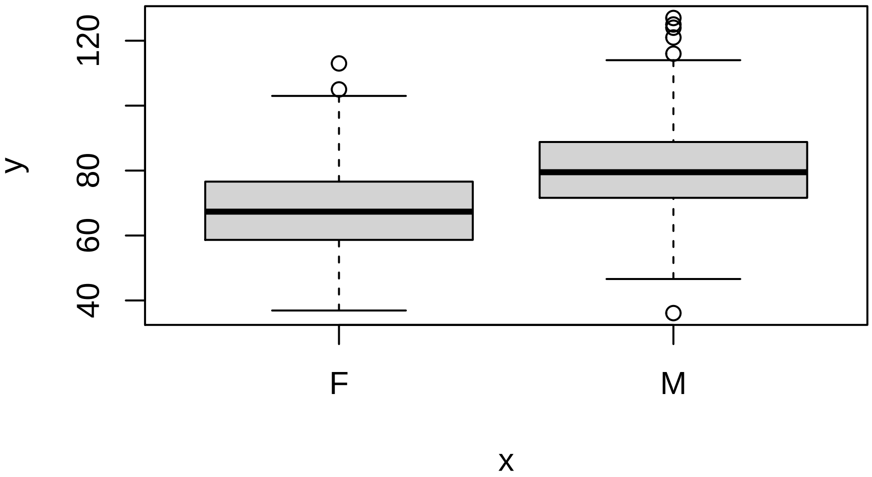
a) We notice that he center of the fixed baseline average is a much lower temperature and teh spread is far less than the distribution for 2005 to 2015. One important thing to note here is that the scale of the two distributions is different: the baseline represents 29 years (1951 to 1980) whereas the distribution in question represents only 10 (2005-2015).

b) The temperatures between the years 2005 and 2015 are likely to be higher on average and fluctuate much more when compared to the temperatures between 1951 and 1980.

Problem 4.

a)

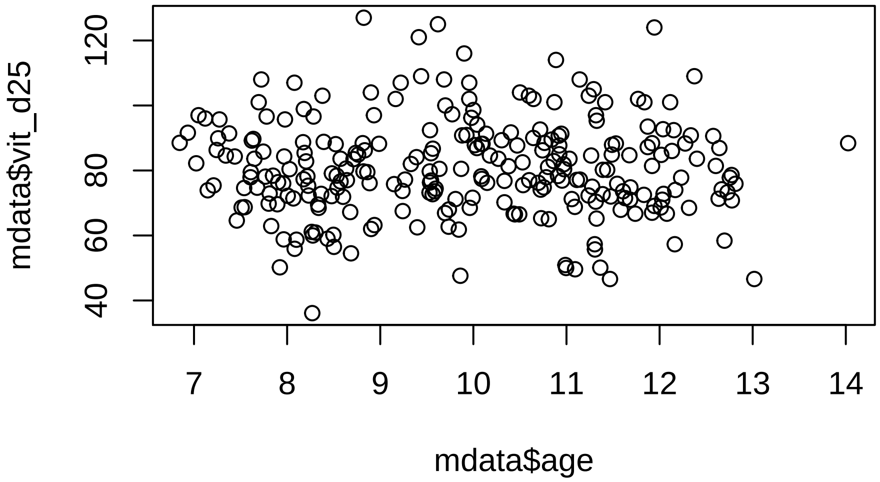
```
#load the data
load("datasets/vitamin_d.Rdata")

#sex and serum 25(OH)D level
plot(vitamin_d$sex, vitamin_d$vit_d25)
```



Interpretation In the boxplots above we see that the median, Q1 and Q3 values for serum 25(OH)D level are higher for Males than Females. The male sub-population also seems to have a larger number of outliers in the high serum level.

b)

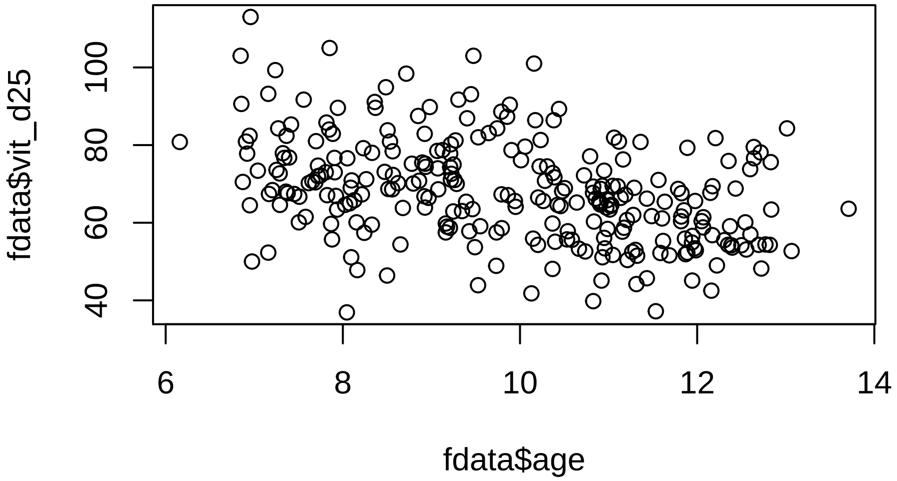
```
#age and serum 25(OH)D level, males
mdata = subset(vitamin_d, vitamin_d$sex == 'M')
plot(mdata$vit_d25 ~ mdata$age)
```



```
cor(mdata$vit_d25, mdata$age, use="complete.obs")
```

```
## [1] -0.008565987
```

```
#age and serum 25(OH)D level, females
fdata = subset(vitamin_d, vitamin_d$sex == 'F')
plot(fdata$vit_d25 ~ fdata$age)
```



```
cor(fdata$vit_d25, fdata$age, use="complete.obs")
```

```
## [1] -0.3976444
```

Interpretation The correlation between serum 25(OH)D level and age is negative for both Males and Females in the sub-population studied, however, the correlation is much higher for females. This indicates that as age increases, the Vitamin D deficiency in Females increases much more drastically in females than in males.

c) One explanation for the stronger negative correlation between serum 25(OH)D level and age among females could be that in this sub-population, the males are probably out in the sun for longer periods of time through out the age-group studied than females, who perhaps, reduce the time spent in the sun as their age increases. In some cultures, unfortunately, males are afforded more opportunity for outdoor sports etc. than females are and this could be one possible reason for this.

d)

```
#Prevalence of vitamin d deficiency
def_data = subset(vitamin_d, vitamin_d$vit_d25 < 50)
#Percentage of children who have deficiency
(nrow(def_data)/nrow(vitamin_d))*100
```

```
## [1] 3.910615
```

In order to calculate the prevalence of Vitamin D deficiency, we can count up the rows with the value of vit_d25 less than 50 as defined by the problem and then calculate the percentag of the total. Based on that it seems like 3.91% of the childeren have the deficiency.

e)

```
#Calculate the number of Males with a Vitamin D deficiency
malesWithDeficiency = nrow(subset(vitamin_d, vitamin_d$vit_d25 < 50 & vitamin_d$sex == "M"))
#Calculate the total number of Males
```

```
totalMales = nrow(subset(vitamin_d, vitamin_d$sex == "M"))
#Calculate proportion of males with deficiency
maleProp = malesWithDeficiency/totalMales
maleProp
```

```
## [1] 0.01879699
```

```
#Calculate the number of Females with a Vitamin D deficiency
femalesWithDeficiency = nrow(subset(vitamin_d, vitamin_d$vit_d25 < 50 & vitamin_d$sex == "F"))
#Calculate the total number of Females
totalFemales = nrow(subset(vitamin_d, vitamin_d$sex == "F"))
#Calculate proportion of females with deficiency
femaleProp = femalesWithDeficiency/totalFemales
femaleProp
```

```
## [1] 0.05904059
```

```
#Calculate Relative Risk: proportion of females with deficiency / proportion of males with def
femaleProp/maleProp
```

```
## [1] 3.140959
```

Interpretation As seen from the calculation above, the relative risk is 3.14, which can be interpreted as the risk for vitamin D deficiency is more than 3 times greater for females than for males within the sub-population studied.

Problem 5.

a)

i.

```
#load the data
load("datasets/stops.Rdata")

#number of stops
nrow(stops)
```

```
## [1] 1756587
```

The police stops are simply calculated by the number of rows in the dataset, since each row represents a stop.

ii.

```
#date range
min(stops$date)
```

```
## [1] "2014-01-01"
```

```
max(stops$date)
```

```
## [1] "2017-12-31"
```

Using the *max* and *min* functions, we can calculate the earliest and last dates.

iii.

```
#proportions of stops occurring in 2017
nrow(subset(stops, date > "2016-12-31" & date < "2018-01-01"))/nrow(stops)
```

```
## [1] 0.2265234
```

Approximately 22% of the stops occured in 2017.

b)

```
#numerical summaries
#Stops for age less than 20
print(paste("Less than 20 yrs: ", nrow(subset(stops, subject_age < 20))))
```

```
## [1] "Less than 20 yrs:  129204"
```

```
#Stops for age between 20 and 30
print(paste("Between 20 & 30 yrs: ", nrow(subset(stops, subject_age > 20 & subject_age < 30)))
```

```
## [1] "Between 20 & 30 yrs:  599990"
```

```
#Stops for age between 30 and 40
print(paste("Between 30 & 40 yrs: ", nrow(subset(stops, subject_age > 30 & subject_age < 40)))
```

```
## [1] "Between 30 & 40 yrs:  356495"
```

```
#Stops for age between 40 and 50
print(paste("Between 40 & 50 yrs: ", nrow(subset(stops, subject_age > 40 & subject_age < 50))))
```

```
## [1] "Between 40 & 50 yrs:  236483"
```

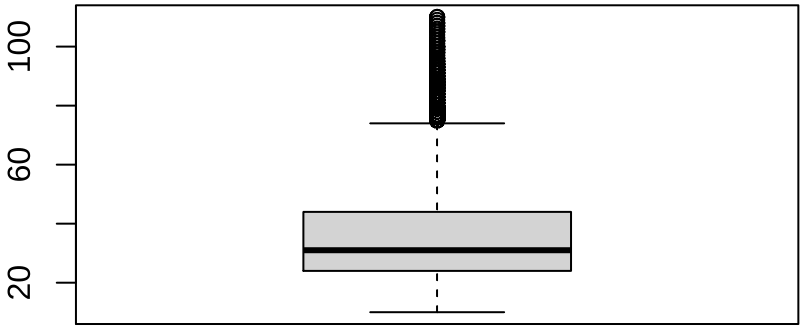
```
#Stops for age between 50 and 65
print(paste("Between 50 & 60 yrs: ", nrow(subset(stops, subject_age > 50 & subject_age < 65))))
```

```
## [1] "Between 50 & 60 yrs:  225030"
```

```
#Stops for age greater than 65
print(paste("More than 65 yrs: ", nrow(subset(stops, subject_age > 65))))
```

```
## [1] "More than 65 yrs:  37719"
```

```
#graphical summaries
boxplot(stops$subject_age)
```



Interpretation We can see-both from the numerical and graphical analysis- that the 20 to 30 years age range was subjected to the maximum number of stops by the police. This would seem to imply that the police seem to suspect persons of this age group as far as possession of contraband.

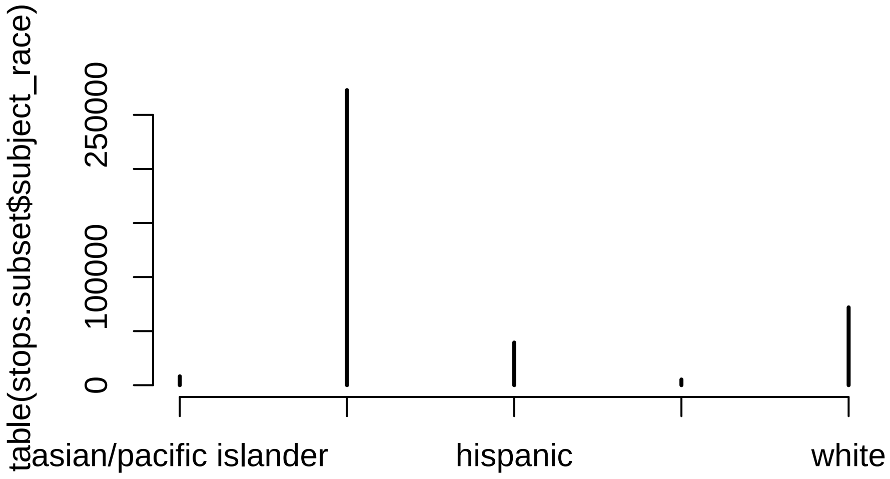
c)

i.

```
#create stops.subset
stops.subset = subset(stops, date > "2016-12-31" & date < "2018-01-01")
#numerical summaries
table(stops.subset$subject_race)
```

```
##
## asian/pacific islander                  black               hispanic
##                   8237                 272946                  39455
##          other/unknown                  white
##                   5275                  71995
```

```
#graphical summaries
plot(table(stops.subset$subject_race))
```



Clearly, the black subset of poeple in the dataset are over-represented.

ii. It is important to understand the racial demographics of Philadelphia, which is the population this sample is drawn from. If the population has a high proportion of black persons, then it is likely that a good sample will be representative of that fact and so the sample of stops will also have more representation from black persons.

iii.

```
#load the data
load("datasets/population_2017.Rdata")
head(population_2017)
```

```
##             subject_race num_people
## 1 asian/pacific islander     110864
## 2                  black     648846
## 3               hispanic     221777
## 4          other/unknown      39858
## 5                  white     548312
```

```
#compute stop rates:



#Stop rate for "asian/pacific islander"
nrow(subset(stops.subset, subject_race == "asian/pacific islander"))/population_2017$num_peopl
```

```
## [1] 0.07429824
```

```
#Stop rate for "black"
nrow(subset(stops.subset, subject_race == "black"))/population_2017$num_people[2]
```

```
## [1] 0.4206638
```

```
#Stop rate for "hispanic"
nrow(subset(stops.subset, subject_race == "hispanic"))/population_2017$num_people[3]
```

```
## [1] 0.1779039
```

```r
#Stop rate for "other/unknown"
nrow(subset(stops.subset, subject_race == "other/unknown"))/population_2017$num_people[4]
```

```
## [1] 0.1323448
```

```r
#Stop rate for "white"
nrow(subset(stops.subset, subject_race == "white"))/population_2017$num_people[5]
```

```
## [1] 0.131303
```

We can see above that the stop rates for Asian/Pacific Hghlander is the lowest at about 7% and that for Black persons is highest at about 42%. There is a huge disparity.

iv.

```r
#Calculating how much more often black drivers are stopped by the police
(nrow(subset(stops.subset, subject_race == "black"))/population_2017$num_people[2])/(nrow(subse
```

```
## [1] 3.203764
```

```r
#Calculating how much more often hispanic drivers are stopped by the police
(nrow(subset(stops.subset, subject_race == "hispanic"))/population_2017$num_people[3])/(nrow(su
```

```
## [1] 1.354912
```

As seen above relative to white drivers, black drivers are stopped 3.2 times more often and hispanic drivers are stopped 1.35 times more often.

d)

i.

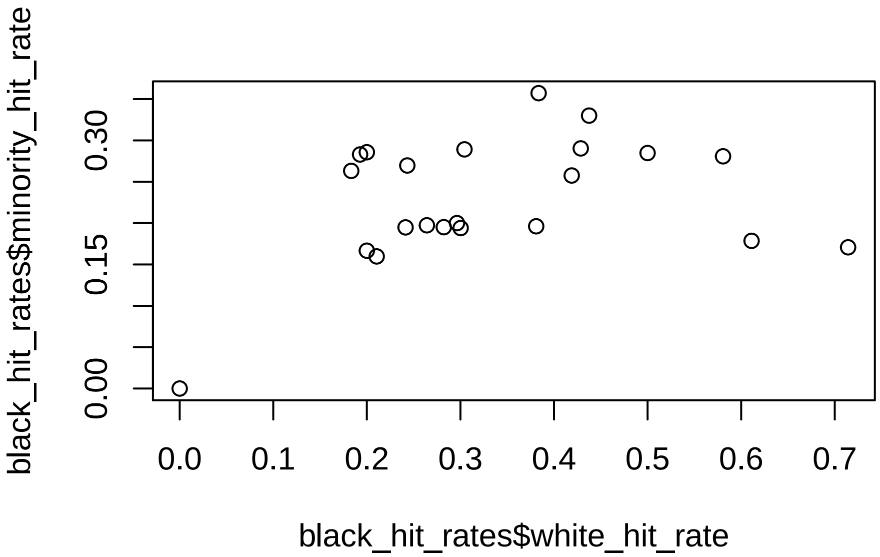
```r
#calculate hit rates
races = list("asian/pacific islander", "black", "hispanic", "other/unknown", "white")
for (x in races)
{
  print(paste("Hit rate for ", x, "is:"))
  print((nrow(subset(stops.subset, subject_race == x & contraband_found == TRUE)))/(nrow(subse
}
```

```
## [1] "Hit rate for  asian/pacific islander is:"
## [1] 0.007405609
## [1] "Hit rate for  black is:"
## [1] 0.01802554
## [1] "Hit rate for  hispanic is:"
## [1] 0.01563807
## [1] "Hit rate for  other/unknown is:"
## [1] 0.007772512
## [1] "Hit rate for  white is:"
## [1] 0.01323703
```

We can see that the hit rate is greatest for black and hispanic drivers at 1.8% and 1.5% with white drivers being close at 1.3%.

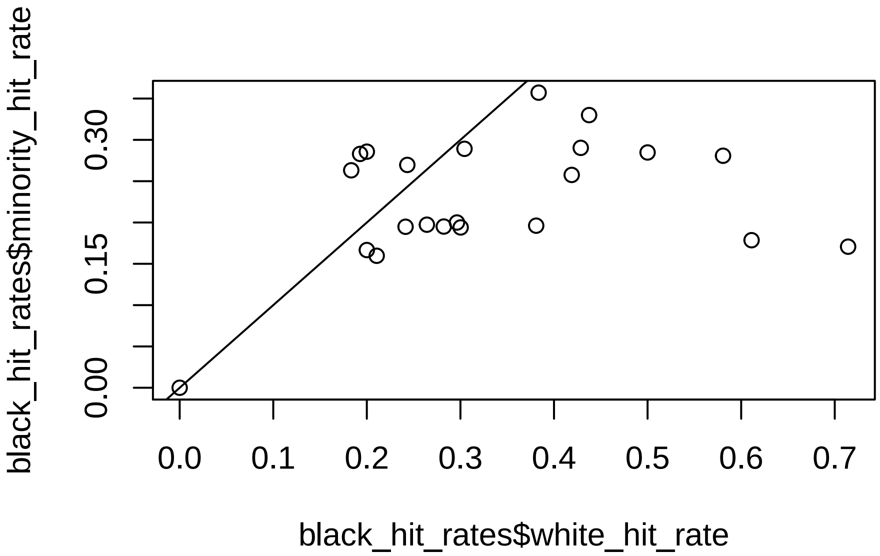
ii.

```r
#load data
load("datasets/hit_rates.Rdata")
#create a plot
black_hit_rates = subset(hit.rates, minority_race == "black")
plot(black_hit_rates$white_hit_rate, black_hit_rates$minority_hit_rate)
```



iii.

```r
#create plot with added y = x line
plot(black_hit_rates$white_hit_rate,
     black_hit_rates$minority_hit_rate)
abline(0,1)
```



Interpretation Clearly, most of the points lie below the y=x line. The line represents points where there the white hit rate is equal to the black hit rate.

iv. From the plot we can see that there is bias against black drivers. Since most of the points fall below the line, it shows us that the hit rate for white drivers is higher that that for black drivers, meaning whn black drivers are stopped, the police find contraband less often than when white drivers are stopped.