Problem Set 1

Kendall Weistroffer & Ariel Huckabay

Due: January 18, 2018

Introduction

The questions below address the data in “ps1_data.csv”. These data are based on 2015 PUMS (Public Use Microdata Sample) data extracted from IPUMS-USA, University of Minnesota, www.ipums.org, with PUMA to county allocations from MABLE, http://mcdc.missouri.edu/websas/geocorr12.html

The rows correspond to sampled families. The first column, “FAMSIZE”, is the number of family members. The second column, “county”, is “Denver” if the family is in a sampling region assigned to Denver County and “Boulder” if the family is in a sampling region assigned to Boulder County. No other counties are represented in these data. The column “FTOTINC” is the total family income. The column “hhwt” gives the number of families represented by the sampled family.If a hhwt is for a family is 33, say, we estimate that 33 families in the study area have the same properties as this family in order to estimate characeristics of the whole population. For example, if you read this data set into a data frame named “dat”, you could calculate sum(dat$FTOTINC*dat$hhwt)/sum(dat$hhwt) to estimate the mean income of all the families.

Please complete the following tasks regarding the data in R. Please generate a solution document in R markdown and upload the .Rmd document and a rendered .doc, .docx, or .pdf document. Your work should be based on the data’s being in the same folder as the .Rmd file. Please turn in your work on Canvas. Your solution document should have your answers to the questions and should display the requested plots.

Questions are 10 points each.

These questions were rendered in R markdown through RStudio (https://www.rstudio.com/wp-content/uploads/2015/02/rmarkdown-cheatsheet.pdf, http://rmarkdown.rstudio.com).

Part 1

Question 1

Please calculate the number of the Denver families that are represented in the data and the number of Boulder families that are represented in the data.

```r
#Read in the dataset:
dat <- read.csv("ps1_data.csv")

#Calculate the number of Denver and Boulder families:
sumDenver <- sum(dat$hhwt[dat$county == "Denver"])
sumBoulder <- sum(dat$hhwt[dat$county == "Boulder"])

#Print:
cat("Denver: ", sumDenver)
```

```
## Denver:  286530.9
```

```r
cat("Boulder: ", sumBoulder)
```

```
## Boulder:  143996.4
```

Question 2

Please estimate the proportion of the Denver families that are 2-person families and the proportion of Boulder families that are two person families. Also provide the over all proportion of the families that are 2-person families.

```
#Grab the total Denver and Boulder 2-Person Families:
denver2Person <- sum(dat$hhwt[dat$county == "Denver"& dat$FAMSIZE ==2])
boulder2Person <- sum(dat$hhwt[dat$county == "Boulder"& dat$FAMSIZE ==2])

#Take their proportions:
denverProp <- denver2Person/sumDenver
boulderProp <- boulder2Person/sumBoulder

#Calculate the total sum and total proportion of two-person families:
totalSum <- sumDenver+sumBoulder
total2Prop <- (denver2Person+boulder2Person)/totalSum



#Print:
cat("Denver Proportion: ", denverProp)
```

```
## Denver Proportion:  0.2986383
```

```
cat("Boulder Proportion: ", boulderProp)
```

```
## Boulder Proportion:  0.3294245
```

```
cat("Total Proportion: ", total2Prop)
```

```
## Total Proportion:  0.3089352
```

Question 3

Please estimate the number of Denver families represented in the data of each size and the number of Boulder families of each size represented in the data.

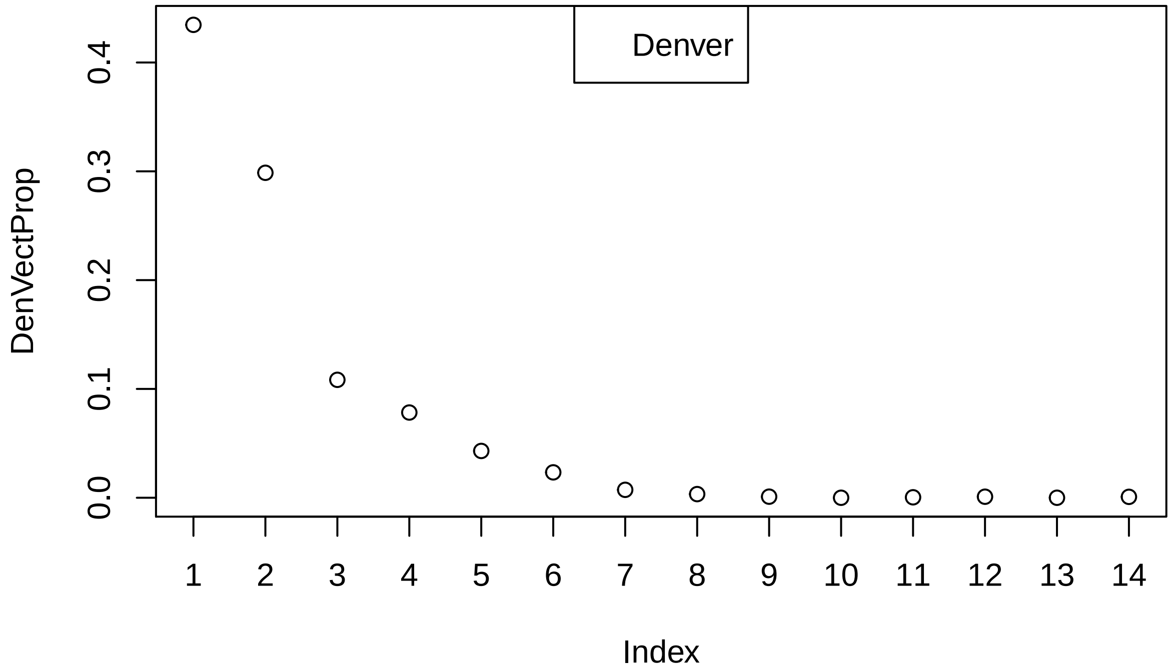
```
#Print out the formatted desired data:
x <- aggregate(hhwt~FAMSIZE+county, data = dat, sum)
print(x)
```

```
##    FAMSIZE  county       hhwt
## 1        1 Boulder  49639.086
## 2        2 Boulder  47435.950
## 3        3 Boulder  18430.656
## 4        4 Boulder  18849.268
## 5        5 Boulder   6876.524
## 6        6 Boulder   2020.518
## 7        7 Boulder    507.250
## 8        8 Boulder    237.188
## 9        1  Denver 124526.430
## 10       2  Denver  85569.090
## 11       3  Denver  31048.430
## 12       4  Denver  22441.310
## 13       5  Denver  12321.290
## 14       6  Denver   6683.060
## 15       7  Denver   2080.250
## 16       8  Denver    957.000
## 17       9  Denver    283.000
## 18      11  Denver    114.000
## 19      12  Denver    269.000
## 20      14  Denver    238.000
```

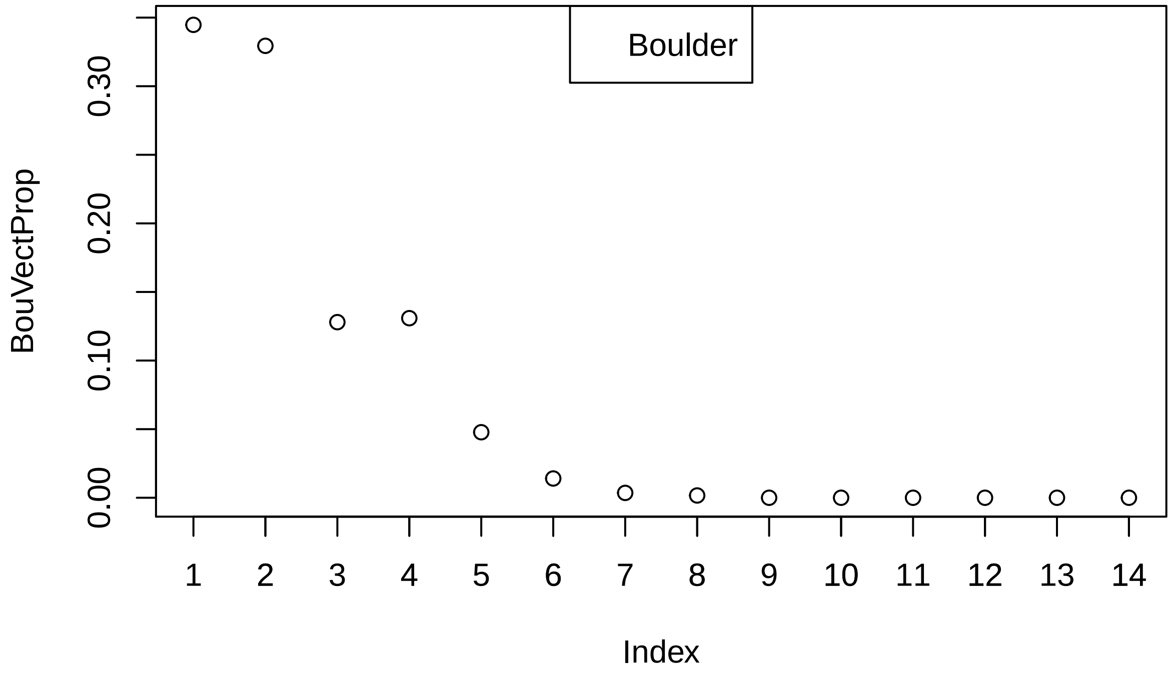
Question 4

Please estimate the proportion of Denver families represented in the data of each size and the proportion of Boulder families of each size represented in the data. Make a plot of the results. You may but are not required to use barplot or ggplot with geom_col(position="dodge") for which partial examples are provided below.

```
DenVector = vector("numeric", max(dat$FAMSIZE))
BouVector = vector("numeric", max(dat$FAMSIZE))

for(i in min(dat$FAMSIZE):max(dat$FAMSIZE)){
  for(j in 1:nrow(dat)){
    if(dat$FAMSIZE[j]==i){
      if(dat$county[j]=="Denver"){DenVector[i]=DenVector[i]+dat$hhwt[j]}
      else{BouVector[i]=BouVector[i]+dat$hhwt[j]}
    }
  }
}

DenVectProp = vector("numeric", max(dat$FAMSIZE))
BouVectProp = vector("numeric", max(dat$FAMSIZE))
for(i in 1:max(dat$FAMSIZE)){
  DenVectProp[i] = DenVector[i]/sumDenver
  BouVectProp[i] = BouVector[i]/sumBoulder
}
plot(DenVectProp, xaxt="n")
legend("top", legend="Denver")
axis(1, at=1:14, labels=1:14)
```



```
plot(BouVectProp)
legend("top", legend="Boulder")
axis(1, at=1:14, labels=1:14)
```

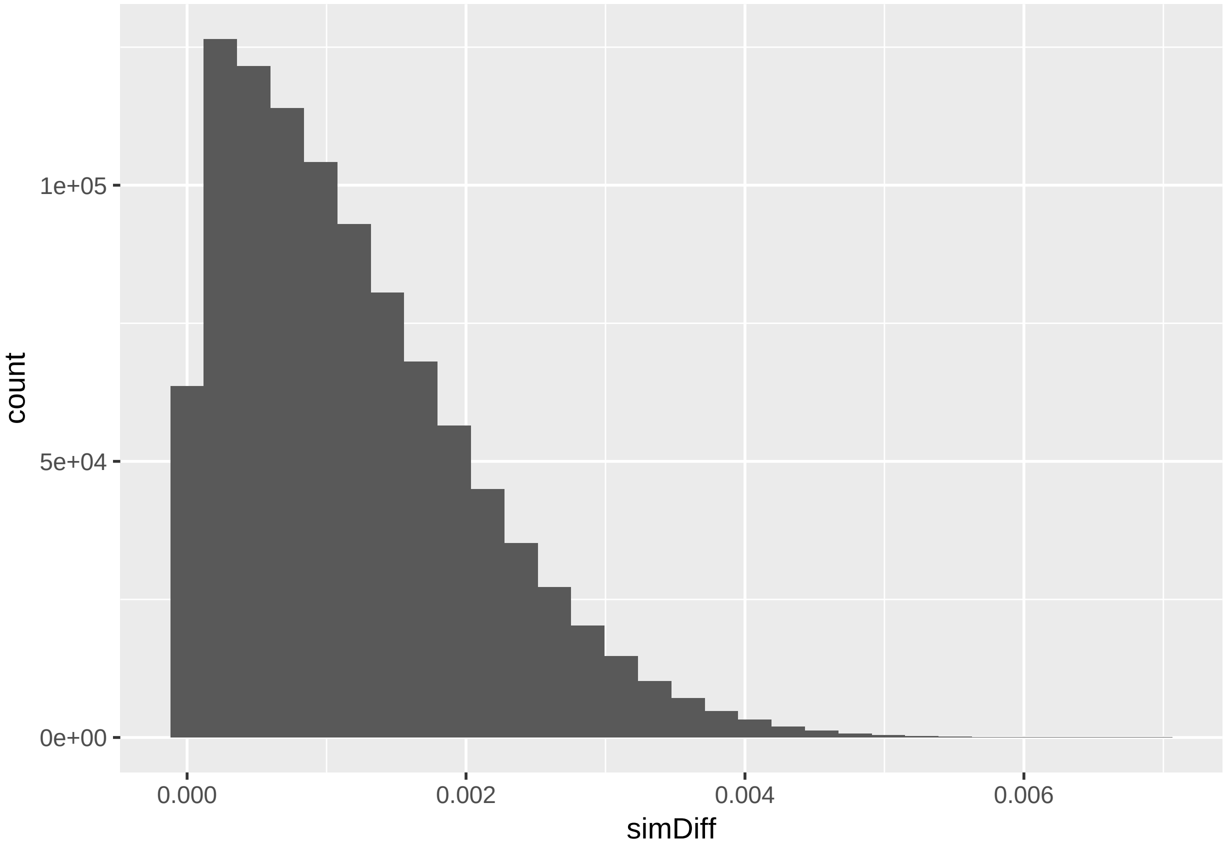


Question 5

Please do a simulation to estimate the probability of obtaining proportions as unequal as this if each family represented in each county is randomly assigned to be a 2-person family with probability equal to 0.3089352. Please use rbinom as in class. You may find the function "round" useful. To emphasize, please use a simulation, not a traditional statistical test.

```
#run 1000000 times
n<-1000000
size<-max(dat$FAMSIZE)

#Run rbinom simulations on Denver and Boulder with the given probability:
simDenver<-rbinom(n, round(sumDenver),0.3089352)
simBoulder <- rbinom(n , round(sumBoulder),0.3089352 )

#Calculate the Proportions:
twosPropD <- simDenver/sumDenver #proportion of Denver 2 person families
twosPropB <- simBoulder/sumBoulder #proportion of Boulder 2 person families

#Take the propotion difference for the simulation and actual data set:
simDiff <- abs(twosPropD-twosPropB)
actDiff <- abs(denverProp - boulderProp)

#Used to show how far off the simulated proportion is from the actual proportion:
qplot(simDiff)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```
#How many simmulated plots are greater than or equal to the actual difference?
output <- mean(actDiff <= simDiff)
cat("The number of simulated plots that are greater than or equal to the actual difference: ", output)
```

```
## The number of simulated plots that are greater than or equal to the actual difference:  0
```

```
#moral of story : our simulated model doesn't match at all the actual data,
#and thus it's very unlikely that this level of inequality would have happened by accident.
```

Question 6 (4441 only)

The estimate in Question 5 randomly assigns family size to each individual family. In the data, there are only 4922 observations. An observation may represent as many as 811 families. How do you think that estimating proportions from a sample, instead of calculating proportions from a census affects the variability of the result? Please devise,carefully explain, and implement a simulation to approximate the probability of obtaining proportions as unequal as observed using a sample like this.

Please note that you don't have enough information to do this perfectly. You will be making a rough estimate. What do you conclude about the consistency of these data with the null hypothesis of no difference between the counties in the proportions of 2-person households?

Some simulations may be slow, so starting with a small number of repetitions can be helpful.

The IPUMS data do include tools to do this fully correctly. Please create a "home made" simulation rather than using these tools. Again, please use a simulation, not a traditional statistical test.

```
n<-100000
nDenverSample<-sum(dat$county=="Denver")
nBoulderSample<-sum(dat$county=="Boulder")

Dens <- dat$county=="Denver"

samDenTwos=0
samBouTwos=0

for (i in 1:nrow(dat)){
  if (Dens[i] == TRUE){
    samDenTwos = samDenTwos + 1
  }else{
    samBouTwos = samBouTwos + 1
  }
}


#Assign families randomly to counties
simDenSample<-rbinom(n, nrow(dat), nDenverSample/nrow(dat))
simBouSample<-rbinom(n, nrow(dat), nBoulderSample/nrow(dat))

#Caluclate mean value from assigning families randomly to counties
propDenSample = mean(simDenSample)
propBouSample = mean(simBouSample)

#Count the number of two person families in the sample and caluclate the proportions
sampTwos = sum(dat$FAMSIZE==2)
propTwos = sampTwos/(nrow(dat))

simDenTwos<-rbinom(n, round(propDenSample), propTwos)
simBouTwos<-rbinom(n, round(propBouSample), propTwos)

#Calculate the mean value from the unweighted sample proportion of twos example
x <- mean(simDenTwos/propDenSample)
y <- mean(simBouTwos/propBouSample)

samDiff = abs(x-y)
actSamDiff = abs(samDenTwos/nrow(dat)-samBouTwos/nrow(dat))
cat("samDiff: ", samDiff)
```

```
## samDiff:  5.956767e-05
```

Repsonse 6:

Estimating from a sample should increase variability in the result versus estimating from census.The strategy for this experiment is to assign each family to a county at random with probabilty equal to the observed proportions to obtain an "accidental" distribution that may more closely resemble a census. Then, we proceed in a manner similar to what was in problem 5 to determine the probability of the observed proportion of 2-person families occurs by accident in an overall population.

The results of this show that while the actual difference is 0.0678586,the experiment has a much more equal distribution haiving a difference of much less (See output on the variable samDiff).