Problem Set 1

C. Durso

January 10, 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.

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.

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.

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.

```
# barplot(dat.sizes$proportion, col=as.factor(dat.sizes$county))
# legend("topright",legend=c("Boulder","Denver"),fill=1:2)
#
# ggplot(data=dat.sizes, aes(x=FAMSIZE,y=proportion,fill=county))+geom_col(position="dodge")
```

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.

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.