Question 4, Homework 1

Joe Froelicher

9/10/2020

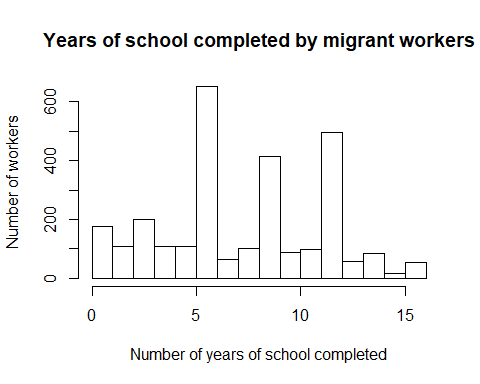
# Part A

setwd("C:/Users/jofro/OneDrive/cu/sem1/methods1/hw1")  
naws <- read.csv("NAWS2014.csv", header = TRUE)  
head(naws$A09)

## [1] 14 6 6 7 12 4

# Part B

hist(  
 naws$A09,  
 main = "Years of school completed by migrant workers",  
 xlab = "Number of years of school completed",  
 ylab = "Number of workers"  
)



# Part C

mean(naws$A09)

## [1] 7.674814

median(naws$A09)

## [1] 7

# https://stackoverflow.com/questions/2547402/how-to-find-the-statistical-mode #  
getMode <- function(v) {  
 uniqv <- unique(v)  
 uniqv[which.max(tabulate(match(v, uniqv)))]  
}  
  
getMode(naws$A09)

## [1] 6

Referencing the histogram, as well as measures of median and mode, we know that data that has a heavy right tail (in this case) or left tail, tend to be better estimated by a median than a mean. It is worth noting that the buckets with the most number of workers are 6, 9, and 12, the years we would expect the most people to drop out of school (For much of the United States those would be the years you are transitioning to a new school).

# Part D

naws["category\_edu"] <- NA  
iter <- 0  
  
for (a in naws$A09) {  
 iter <- iter + 1  
   
 if (a <= 5) {  
 naws$category\_edu[iter] <- "00-05"  
   
 } else if (a <= 8) {  
 naws$category\_edu[iter] <- "06-08"  
   
 } else if (a <= 11) {  
 naws$category\_edu[iter] <- "09-11"  
   
 } else {  
 naws$category\_edu[iter] <- "12+"  
 }  
}  
  
head(naws$category\_edu)

## [1] "12+" "06-08" "06-08" "06-08" "12+" "00-05"

# Part E

cat\_table <- table(naws$category\_edu) / length(naws$category\_edu)  
cat\_table

##   
## 00-05 06-08 09-11 12+   
## 0.2479632 0.2890542 0.2118314 0.2511513

# Part F

set.seed(8675309)  
mockdata <- data.frame(subject = 1:800, random = runif(n = 800))  
head(mockdata)

## subject random  
## 1 1 0.1594836  
## 2 2 0.4781883  
## 3 3 0.7647987  
## 4 4 0.7696877  
## 5 5 0.2685485  
## 6 6 0.6730459

mockdata["educ\_cat"] <- NA  
mockdata["educ\_years"] <- NA  
mockdata["educ\_stop\_yn"] <- NA  
  
level1 <- cat\_table[1]  
level2 <- cat\_table[1] + cat\_table[2]  
level3 <- cat\_table[1] + cat\_table[2] + cat\_table[3]  
iter <- 0  
  
for (r in mockdata$random) {  
 iter <- iter + 1  
   
 if (r <= level1) {  
 mockdata$educ\_cat[iter] <- "00-05"  
 mockdata$educ\_years[iter] <- runif(1, 0, 6)  
 mockdata$educ\_stop\_yn[iter] <- rbinom(1, 1, 0.8)  
   
 } else if ( r > level1 & r <= level2 ) {  
 mockdata$educ\_cat[iter] <- "06-08"  
 mockdata$educ\_years[iter] <- runif(1, 6, 9)  
 mockdata$educ\_stop\_yn[iter] <- rbinom(1, 1, 0.8)  
   
 } else if ( r > level2 & r <= level3 ) {  
 mockdata$educ\_cat[iter] <- "09-11"  
 mockdata$educ\_years[iter] <- runif(1, 9, 12)  
 mockdata$educ\_stop\_yn[iter] <- rbinom(1, 1, 0.8)  
   
 } else {  
 mockdata$educ\_cat[iter] <- "12+"  
 mockdata$educ\_years[iter] <- runif(1, 12, 17)  
 mockdata$educ\_stop\_yn[iter] <- rbinom(1, 1, 0.8)  
   
 }  
}  
  
iter <- 0  
for (c in mockdata$educ\_cat) {  
 iter <- iter + 1  
   
 if ( c == "06-08" & mockdata$educ\_stop\_yn[iter] == 1 ) {  
 mockdata$educ\_years[iter] <- 6  
 } else if ( c == "09-11" & mockdata$educ\_stop\_yn[iter] == 1 ) {  
 mockdata$educ\_years[iter] <- 9  
 } else if ( c == "12+" & mockdata$educ\_stop\_yn[iter] == 1 ) {  
 mockdata$educ\_years[iter] <- 12  
 }  
}  
  
mockdata$educ\_years <- trunc(mockdata$educ\_years)  
head(mockdata)

## subject random educ\_cat educ\_years educ\_stop\_yn  
## 1 1 0.1594836 00-05 3 1  
## 2 2 0.4781883 06-08 6 1  
## 3 3 0.7647987 12+ 12 1  
## 4 4 0.7696877 12+ 12 1  
## 5 5 0.2685485 06-08 6 1  
## 6 6 0.6730459 09-11 9 1

# Part G

hist(  
 mockdata$educ\_years,  
 main = "Years of school completed (mockdata)",  
 xlab = "Number of years of school completed",  
 ylab = "Number of workers",  
 breaks = 16  
)

