#컴퓨터 재무학 과제1. 20162878 경제학과 김홍식

#C1, file : wage1

#(i) Find the average education level in the sample. What are the lowest and highest years of education?

mean(data$educ) #평균 12.562

max(data$educ) #최대 18

min(data$educ) #최소 0

#(ii) Find the average hourly wage in the sample. Does it seem high or low?

mean(data$wage) #평균 5.896

median(data$wage) #중간값 4.65

hist(data$wage) #중간값이 4.65이고 히스토그램이 우측으로 긴 꼬리를 가지므로 낮다고 판단됨

#C2, file : bwght

#(i) How many women are in the sample, and how many report smoking during pregnancy?

dim(data) #1388명의 샘플 존재

length(data$cigs[data$cigs>0]) #212명이 임신중 흡연

#(ii) What is the average number of cigarettes smoked per day? Is the average a good measure of the “typical” woman in this case? Explain.

mean(data$cigs) #평균 2.08개비 흡연, 그러나 대부분의 여성은 흡연을 하지 않으므로 대표성을 갖지 못함

#(iii) Among women who smoked during pregnancy, what is the average number of cigarettes smoked per day? How does this compare with your answer from part (ii), and why?

mean(data$cigs[data$cigs>0]) #흡연자들은 평균 13.665개비를 소비하고 전체 여성은 평균 2.08개비를 흡연함. 의미있는 분석을 위해선 비흡연자와 흡연자를 구분해야 함

#(iv) Find the average of fatheduc in the sample. Why are only 1,192 observations used to compute this average?

length(na.omit(data$fatheduc)) #결측값을 제외한 유의미한 데이터의 수가 1192뿐이다

mean(na.omit(data$fatheduc)) #결측값을 제외한 데이터의 평균값은 13.186이다.

#(v) Report the average family income and its standard deviation in dollars.

mean(data$faminc) #평균 29.026

sd(data$faminc) # 표준편차 18.73928

#C6, file : countymurders, 1996년 자료만 이용

#(i) How many counties are there in the data set? Of these, how many have zero murders? What percentage of counties have zero executions? (Remember, use only the 1996 data.)

data1996 <- data[data$year == 1996,]

length(unique(data$countyid)) #카운티 갯수 2197개

zero <- nrow(subset(data1996, data1996$execs ==0)) #살인이 없었던 county 2166개

inte <- nrow(subset(data1996, data1996$execs >0)) #살인이 있었던 county 31개

zero / 2197 #31/2197 = 0.98588, 98.5%의 county는 살인이 없었음

#(ii) What is the largest number of murders? What is the largest number of executions? Compute the average number of executions and explain why it is so small.

max(data1996$murders) #가장 많은 살인자 수 1403

max(data1996$execs) #가장 많은 사형집행 수 3

mean(data1996$execs) #평균 0.015 명이 사형을 집행당함

hist(data1996$execs) #대부분의 county가 사형을 집행하지 않았음

#(iii) Compute the correlation coefficient between murders and execs and describe what you find.

cor(data1996$execs, data1996$murders) #상관관계 0.209 약 20%, 약한 상관관계를 가짐

#(iv) You should have computed a positive correlation in part (iii). Do you think that more executions

#cause more murders to occur? What might explain the positive correlation?

cor(data1996$density, data1996$murders) #상관관계 0.353 약 35%, 사형집행보다 더 높은 양의 상관관계를 보임.