ADV.STATS.PROJECT.R

hadeelmusallam

2022-12-04

#Abstract  
#There has been many times where others have asked whether or not there is a difference in income for Americans   
#with different races, ages, marital status, or sex. For this project, I will be using the census dataset from   
#the package ‘openintro’ package. The purpose of this project will be to determine whether or not these biases   
#in status have an effect on an American’s income. If there are, we will determine which factors correlate with   
#the hypothesis and how strong/weak the relationships are. The data is a random sample for the 2000 U.S. Census   
#Data in Open Intro. There are 500 observations on the following 8 variables: census year, name of state,   
#total family income, age of each participant, the race of the participants (Black, White, Chinese, American   
#Indian, Alaska Native, Japanese, and Other), the marital status of participants (Divorced, Married/Spouse   
#Absent, Married/Spouse Present, Never Married/Single, Separated and Widowed), and the total personal income of   
#the participant.   
  
#Load Openintro library  
install.packages("openintro",repos='http://cran.us.r-project.org')

##   
## The downloaded binary packages are in  
## /var/folders/21/1whlzzgd6p7bgc4rz9d1jlp40000gn/T//RtmpO8dilf/downloaded\_packages

library(openintro)

## Loading required package: airports

## Loading required package: cherryblossom

## Loading required package: usdata

#Analyze the census dataset  
str(census)

## tibble [500 × 8] (S3: tbl\_df/tbl/data.frame)  
## $ census\_year : int [1:500] 2000 2000 2000 2000 2000 2000 2000 2000 2000 2000 ...  
## $ state\_fips\_code : Factor w/ 47 levels "Alabama","Arizona",..: 9 9 9 9 9 9 9 9 9 9 ...  
## $ total\_family\_income : int [1:500] 14550 22800 0 23000 48000 74000 23000 74000 60000 14600 ...  
## $ age : int [1:500] 44 20 20 6 55 43 60 47 54 58 ...  
## $ sex : Factor w/ 2 levels "Female","Male": 2 1 2 1 2 1 1 1 1 1 ...  
## $ race\_general : Factor w/ 8 levels "American Indian or Alaska Native",..: 7 8 2 8 8 8 8 8 2 8 ...  
## $ marital\_status : Factor w/ 6 levels "Divorced","Married/spouse absent",..: 3 4 4 4 3 3 3 3 3 6 ...  
## $ total\_personal\_income: int [1:500] 0 13000 20000 NA 36000 27000 11800 48000 40000 14600 ...

head(census)

## # A tibble: 6 × 8  
## census\_year state\_fips\_code total\_family\_income age sex race\_general   
## <int> <fct> <int> <int> <fct> <fct>   
## 1 2000 Florida 14550 44 Male Two major races  
## 2 2000 Florida 22800 20 Female White   
## 3 2000 Florida 0 20 Male Black   
## 4 2000 Florida 23000 6 Female White   
## 5 2000 Florida 48000 55 Male White   
## 6 2000 Florida 74000 43 Female White   
## # … with 2 more variables: marital\_status <fct>, total\_personal\_income <int>

summary(census)

## census\_year state\_fips\_code total\_family\_income age   
## Min. :2000 California : 62 Min. : 0 Min. : 0.0   
## 1st Qu.:2000 New York : 43 1st Qu.: 21500 1st Qu.:17.0   
## Median :2000 Florida : 39 Median : 43000 Median :35.0   
## Mean :2000 Texas : 37 Mean : 57411 Mean :35.3   
## 3rd Qu.:2000 Pennsylvania: 26 3rd Qu.: 70700 3rd Qu.:51.0   
## Max. :2000 Ohio : 23 Max. :892050 Max. :93.0   
## (Other) :270 NA's :15   
## sex race\_general  
## Female:232 White :363   
## Male :268 Black : 71   
## Other : 25   
## Two major races : 16   
## Other Asian or Pacific Islander: 13   
## Chinese : 8   
## (Other) : 4   
## marital\_status total\_personal\_income  
## Divorced : 38 Min. : -4400   
## Married/spouse absent : 14 1st Qu.: 5900   
## Married/spouse present:192 Median : 17750   
## Never married/single :222 Mean : 29082   
## Separated : 3 3rd Qu.: 37000   
## Widowed : 31 Max. :456000   
## NA's :108

#Clean the dataset.   
#Analizing the total family income  
children <- nrow(census[census$age < 15, ])  
  
#Remove the 108 Children from the data set since they do not have a total personnal income listed   
Income <- subset(census, !is.na(census$total\_personal\_income))  
  
#Analysis. Explore the race, age, martial status, and sex variables of the dataset.   
summary(census$race\_general)

## American Indian or Alaska Native Black   
## 3 71   
## Chinese Japanese   
## 8 1   
## Other Other Asian or Pacific Islander   
## 25 13   
## Two major races White   
## 16 363

summary(census$age)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.0 17.0 35.0 35.3 51.0 93.0

summary(census$marital\_status)

## Divorced Married/spouse absent Married/spouse present   
## 38 14 192   
## Never married/single Separated Widowed   
## 222 3 31

summary(census$sex)

## Female Male   
## 232 268

#Difference in Income based on race.  
#As seen in the exploratory analysis, the two major races are Black and White nationalities.   
#Other races represent a small fraction of the total sample size, making their analyses less significant.   
#We will determine the difference in income between the Black and White population, by running a two-independent   
#sample t-test.  
#Hypotheses  
#μ1= Black  
#μ2= White  
#H0:μ1−μ2=0  
#HA:μ1−μ2≠0  
#Two-tailed test  
  
#Black subset  
Black <- subset(census, census$race\_general == "Black")  
  
#White subset  
White <- subset(census, census$race\_general == "White")  
  
#Mean difference in total personal income (x1 - x2)  
mean(Black$total\_personal\_income, na.rm = TRUE) - mean(White$total\_personal\_income, na.rm = TRUE)

## [1] -15171.02

# T-test function  
t.test(Black$total\_personal\_income, White$total\_personal\_income, alternative = "two.sided")

##   
## Welch Two Sample t-test  
##   
## data: Black$total\_personal\_income and White$total\_personal\_income  
## t = -4.019, df = 231.87, p-value = 7.903e-05  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -22608.300 -7733.746  
## sample estimates:  
## mean of x mean of y   
## 17049.19 32220.22

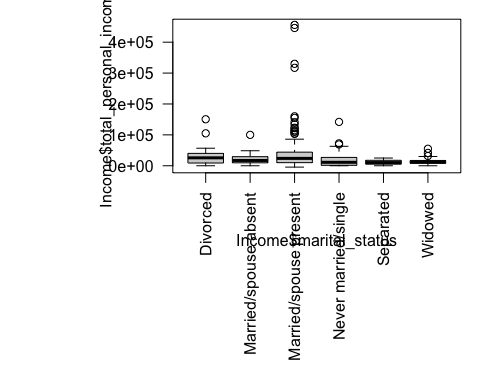
#Decision  
#p > 0.05, therefore we fail to reject Ho.  
  
#Conclusion  
#There is sufficient evidence to support the claim that there is a significant difference in   
#income between the Black and White races.  
  
#Difference in income based on Age  
#scatterplot of the relationship between income and age  
plot(Income$age, Income$total\_personal\_income, main = "Total Personal Income per Age",  
 xlab = "Income", ylab = "Age")



#based on the plot, there is a weak relationship between total personal income and age.   
  
#calculate the coefficient  
cor(Income$age, Income$total\_personal\_income)

## [1] 0.1314763

#The correlation coefficient is very little, this implies that age does not have a significant impact on   
#income.   
  
#Difference in Income based on Martial status.   
#This stage of the analysis will use the analysis of variance(ANOVA) to see if there is a difference  
#between income and martial status.   
  
#Structure a boxplot for the relationship between martial status and personal income.   
par(mar=c(11,9,1,1))  
boxplot(Income$total\_personal\_income~Income$marital\_status, las= 2)



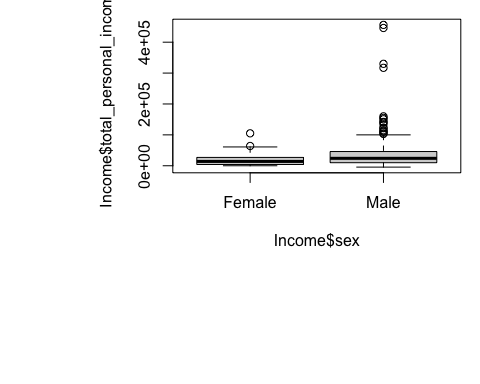
#This side-by-syde scatterplot of the total personal income for the 392 Americans in   
#each marital status group (divorced, married/spouse present, married/spouse absent, widowed,   
#separated, never married/single) shows quite a bit of details (e.g.,the outliers in the married/spouse   
#present group).  
  
#Preform analysis variance.   
#Hypothesis  
#H0=μ1=μ2=μ3=μ4=μ5=μ6  
#HA:At least one mean is different.  
  
#Determine the p-value and the test statistics.   
#use aov() and store as "result"  
result<- aov(Income$total\_personal\_income~Income$marital\_status)  
  
#to find the p-value, use the summary function.   
summary(result)

## Df Sum Sq Mean Sq F value Pr(>F)   
## Income$marital\_status 5 4.145e+10 8.290e+09 4.055 0.00134 \*\*  
## Residuals 386 7.891e+11 2.044e+09   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

#Decision and Conclusion  
#Since p < 0.05 we reject Ho, and the data presents sufficient   
#evidence to support the claim that the income for all groups   
#of different marital status differs.  
#The results above shows that the F-Statistics is 4.06, although it   
#can be used to determine whether or not there are significant differences   
#in income between the different groups, it does not show how they differ.   
#So further analysis is needed.  
  
#we will determine how the personal income differs between the different marital statuses   
#using the paiwise.t.test function.  
pairwise.t.test(Income$total\_personal\_income,Income$marital\_status,  
 p.adjust.method = "bonferroni", na.rm = TRUE)

##   
## Pairwise comparisons using t tests with pooled SD   
##   
## data: Income$total\_personal\_income and Income$marital\_status   
##   
## Divorced Married/spouse absent Married/spouse present  
## Married/spouse absent 1.0000 - -   
## Married/spouse present 1.0000 1.0000 -   
## Never married/single 1.0000 1.0000 0.0013   
## Separated 1.0000 1.0000 1.0000   
## Widowed 1.0000 1.0000 0.0737   
## Never married/single Separated  
## Married/spouse absent - -   
## Married/spouse present - -   
## Never married/single - -   
## Separated 1.0000 -   
## Widowed 1.0000 1.0000   
##   
## P value adjustment method: bonferroni

#Based on the above analysis, on a scale of 1.00, there are significant differences   
#in income between the Never married/single and the Married/spouse present groups.   
#The results suggest that there is a difference in income based on marital status of Americans,   
#especially if Married/spouse present or Never married/single.  
  
#Difference in Income based on Sex  
#The sex variable has two different categories : Males and Females. Determine the difference   
#in income, if there is any, between these two categories.  
#plot the relationship betweem sex and personal income.   
boxplot(Income$total\_personal\_income~Income$sex)



#This scatterplot shows that there are a few outliers present, but because there the sample   
#size is large enough, these outliers are not of major concern.  
#Hypotheses H0:μfemale=μmale  
#HA:μfemale≠μmale  
#We will run t.test to determine the p-value to see which whether or not   
#Ho will be rejected.   
  
#Create Males and Females subsets   
maleincome <- subset(Income,sex=="Male")  
femaleincome <- subset(Income,sex=="Female")  
  
#T.test  
t.test(femaleincome$total\_personal\_income,maleincome$total\_personal\_income)

##   
## Welch Two Sample t-test  
##   
## data: femaleincome$total\_personal\_income and maleincome$total\_personal\_income  
## t = -5.118, df = 237.12, p-value = 6.38e-07  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -30977.56 -13758.03  
## sample estimates:  
## mean of x mean of y   
## 17441.34 39809.14

#Decision and Conclusion  
#Since p < 0.5, we reject Ho. The data provides sufficient evidence to support the claim  
#that there is significant difference in total personal income between males and females  
  
  
#Conclusions  
#All the analyses conducted above were to answer the question of whether or not there   
#is a difference in American’s income based on sex, age, marital status, and race.   
#To answer that question, we used both Two-independent and Non-independent Sample T-tests,   
#a Multiple Comparison T-test along with an ANOVA(multiple means) Test, and Linear Regression.   
#The results were as follow for each variable analyzed.  
  
#Race   
#There is significant difference in income between the Black and White races,   
#with the white population earning a higher personal income than the black population.  
  
#Age   
#Children (<15 years-old) were excluded from the data, since their income was not applicable.   
#The sample size analyzed was 392, and age did not appear to be a determinant factor of differences in personal income.  
  
#Marital Status   
#The only groups among which there existed a significant difference in income were:   
#the Never married/single and the Married/spouse present groups.  
  
#Sex Analyses   
#It shows there is a significant difference in income between males and females, with males   
#having a higher personal income than females.