**[Question no. 1]** R has been installed

**[Question no. 2]** GSS.DATA <- read.csv("GSSdata.txt", header = TRUE, sep = ",", dec = ".")

**[Question no. 3]** Cleaning Script loaded and run (selecting all and pressing run)

**[Question no. 4]** Plan and conduct two different t-tests. For each one:

# **FIRST SET OF VARIABLES: Total Family Income & Race**

**[Performing T-tests] Relationship between total family income and race:**

Part (a & b) The racial wealth gap between white and non-white (particularly African American) residents in the United States has been the subject of study by economists. According to the 2016 Survey of Consumer Finances the average wealth of households with a head identifying as black was $140,000, while the corresponding level for white-headed households was $901,000, nearly 6.5 times greater (Aliprantis, 2019). Therefore, we expect to find a strong relationship between total family income and race. The prediction is that we will see an unequal mean income amongst white and non-white us residents.

(Part c)

**Null hypothesis:** The difference between the mean income of white US residents and non-white US residents is 0.

**Alternative hypothesis:** The difference between the mean income of white US residents and non-white US residents is not 0.

(Part d):

**Code:**

# RACE already cleaned

# Cleaning INCOME

GSS.data$INCOME[GSS.data$INCOME==0 | GSS.data$INCOME >= 13] <- NA

# Creating column WHITE (1 if WHITE else 0)

GSS.data$WHITE <- ifelse(GSS.data$RACE==1, 1, 0)

GSS.data$WHITE <- as.factor(GSS.data$WHITE)

# Exploratory data analysis

Summary(GSS.data$WHITE)

NON-WHITE WHITE

10711 46350

# Performing t-test

t.test(INCOME~WHITE, data=GSS.data)

**Output:**

Welch Two Sample t-test

data: INCOME by WHITE

t = -24.84, df = 12026, p-value < 2.2e-16

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-1.0076647 -0.8602641

sample estimates:

mean in group 0 mean in group 1

9.187759 10.121723

**Code:**

model <- t.test(INCOME~WHITE, data=GSS.data)

model

#We can also get just the confidence intervals by running

model$conf.int

#For effect size

#We use the library

library('effsize')

cohen.d(INCOME~WHITE, data=GSS.data, mu=2)

**Output:**

Cohen's d

d estimate: -0.3276507 (small)

95 percent confidence interval:

lower upper

-0.3504082 -0.3048932

(Part e) The 95% confidence intervals are from -1.0076647 to -0.8602641, meaning we are 95% confident that the difference between of mean for Non-White and White respondents was between -1.0076647 to -0.8602641. The effect size is -0.3276507 small, which means its not negligible but further probing and increasing sample size might be considered. With P size 16 decimal places it can be stated that the chances are almost 100% that the difference between White and Nonwhites income is not zero. The obtained p-value 2.2e-16 is far smaller than the commonly used significance level of 0.05. It indicates strong evidence against the null hypothesis, as there is less than a 5% probability the null is correct (and the results are random). Comparing means between both groups is an important method for identifying discrimination and other social problems in this case the sample mean is 9.187759 10.121723 evident of the discrimination / social problem being prevalent. Therefore, we reject the null hypothesis, and accept the alternative hypothesis. Our analysis shows that there is a relationship between income and race, supporting the results of earlier works on the topic. Hence, our analysis agrees with our prediction.

# **SECOND SET OF VARIABLES: Respondents Degree & Father being a college Graduate**

Relationship between respondent’s degree and respondent’s father being a college graduate:

Parents are the primary influence in their children’s lives. Research shows that parents’ education level has a significant impact on their children’s success. A 2014 College Board/National Journal survey exploring Americans’ educational choices found that 80 percent of those raised by two graduates said their parents encouraged them to attend a four-year school, compared with 29 percent of those raised in families without a degree (Brownstein, 2014). We therefore expect to find a strong relationship between the respondent’s degree and the respondent’s father being a college graduate.

**Null Hypothesis:** Respondents degree as a result of a degree holding father is unrelated

**Alternative Hypothesis:** Respondents earning a degree and their father having a degree is related

# Creating column FTHCLG (1 if respondent's father graduated from college, 0 # otherwise). Does not include junior college.

GSS.data$FTHCLG <- ifelse((GSS.data$PADEG == 3 | GSS.data$PADEG == 4), 1, 0)

GSS.data$FTHCLG <- as.factor(GSS.data$FTHCLG)

# Cleaning column DEGREE

GSS.data$DEGREE[GSS.data$DEGREE >= 7] <- NA

# t-test

t.test(DEGREE~FTHCLG,data=GSS.data)

**Output:**

Welch Two Sample t-test

data: DEGREE by FTHCLG

t = -74.081, df = 7233, p-value < 2.2e-16

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-1.238977 -1.175097

sample estimates:

mean in group 0 mean in group 1

1.188783 2.395820

**Code:**

Model2 <- t.test(DEGREE~FTHCLG, data=GSS.data)

Model2

#We can also get just the confidence intervals by running

Model2$conf.int

#For effect size

#We use the library

library('effsize')

cohen.d(DEGREE~FTHCLG, data=GSS.data, mu=2)

**Output:**

Cohen's d

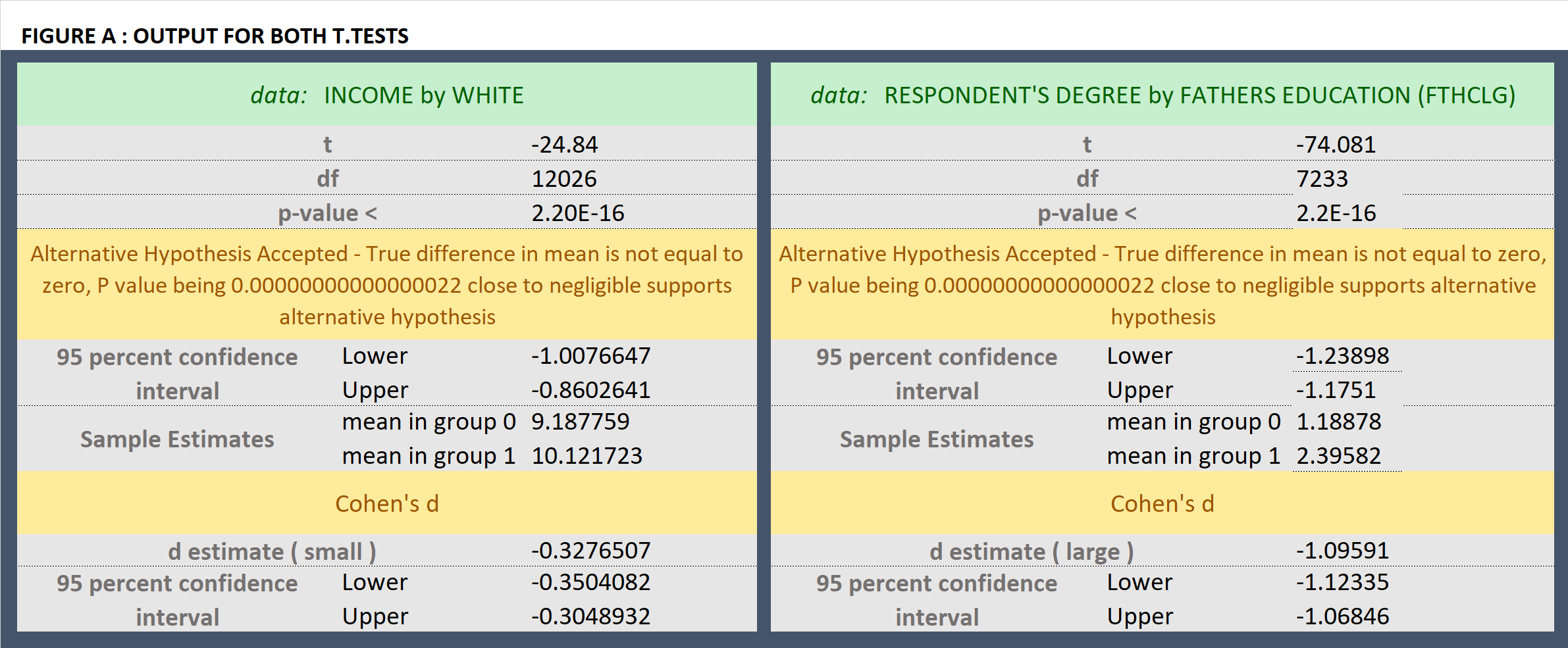
d estimate: -1.095907 (large)

95 percent confidence interval:

lower upper

-1.123354 -1.068459

The 95% confidence interval is from -1.238977 to -1.175097. The effect size is -1.095907 large. The obtained p-value 2.2e-16 is far smaller than the commonly used significance level of 0.05. It indicates strong evidence against the null hypothesis; Therefore, we reject the null hypothesis, and accept the alternative hypothesis. Our analysis shows that there is a relationship between the highest degree obtained by a respondent and the respondent’s father being a college graduate. Hence, our analysis agrees with our prediction.



**[Question no. 5] Correlation Matrix Code**

[[1]](#footnote-1)A correlation matrix is a tabulation of correlation coefficient for the variables in a set of data which show their relationship. The coefficient indicates both the strength of the relationship as well as the direction. In R it is very convenient to learn more about the data by using the plot function to graphically represent the data and to use the corr command to see the correlations.

We install the following package:

install.packages("Hmisc")

library("Hmisc")

We use the following code to run the correlation matrix with p-values. The data has to be fed to the rcorr function as a matrix.

mydata <- GSS.data[c('Variable 1', 'Variable 2'," 'Variable 3","Variable 4")]

mydata.rcorr = rcorr(as.matrix(mydata))

mydata.rcorr

This generates one table of correlation coefficients (the correlation matrix) and another table of the p-values. To extract the values from this object into a useable data structure, you can use the following syntax:

For correlation coefficients or correlation matrix

mydata.coeff = mydata.rcorr$r

For p-values

mydata.p = mydata.rcorr$P

**[Question no. 6] Correlation Matrix and P-values of Four Selected Variables**

Install packages:

install.packages("Hmisc")

library("Hmisc")

Code for cleaning data (RACE and JOBSECOK were already cleaned)

GSS.data$RINCOME[GSS.data$RINCOME==0 | GSS.data$RINCOME > 12] <- NA

GSS.data$CLASS[GSS.data$CLASS==0 | GSS.data$CLASS > 4] <- NA

(Part a) Selecting the four variables of our interest:

mydata <- GSS.data[c('RACE','CLASS',"RINCOME","JOBSECOK")]

mydata

mydata.rcorr = rcorr(as.matrix(mydata))

(Part b) For correlation coefficients

mydata.rcorr$r

Creating Table 1[[2]](#footnote-2):

library(apaTables)

Using Package WriteXl

apa.cor.table(mydata)

apa.cor.table(mydata, filename="table1.doc")

Table 1

*Means, standard deviations, and correlations with confidence intervals*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variable | *M* | *SD* | 1 | 2 | 3 |
|  |  |  |  |  |  |
| 1. RACE | 1.24 | 0.53 |  |  |  |
|  |  |  |  |  |  |
| 2. CLASS | 2.46 | 0.66 | -.12\*\* |  |  |
|  |  |  | [-.13, -.11] |  |  |
|  |  |  |  |  |  |
| 3. RINCOME | 9.09 | 3.40 | -.01\* | .16\*\* |  |
|  |  |  | [-.02, -.00] | [.15, .17] |  |
|  |  |  |  |  |  |
| 4. JOBSECOK | 1.67 | 0.86 | .04\*\* | -.09\*\* | -.06\*\* |
|  |  |  | [.02, .07] | [-.12, -.07] | [-.09, -.04] |
|  |  |  |  |  |  |

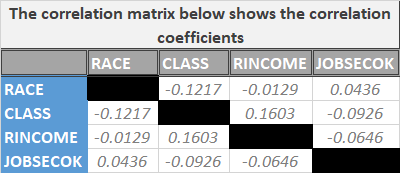
*Note.* *M* and *SD* are used to represent mean and standard deviation, respectively. Values in square brackets indicate the 95% confidence interval for each correlation. The confidence interval is a plausible range of population correlations that could have caused the sample correlation (Cumming, 2014). \* indicates *p* < .05. \*\* indicates *p* < .01.

**Correlation Matrix:**

We choose four variables (race, class, respondent’s income and job security) and we try to observe the correlation between these variables[[3]](#footnote-3).

(Part b,c) For correlation coefficients

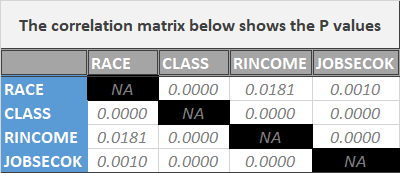
mydata.rcorr$r



(Part b,c) For p-values

mydata.rcorr$P

**P-values Matrix:**

****

(Part d) The two tables show the correlationsbetween the four selected variables. P-values of < 0.05, which is the commonly used significance level, denote that the null hypothesis should be rejected in favor of the alternative hypothesis. In this scenario, the null hypothesis is that there is no correlation between two variables (i.e. correlation coefficient is 0). From the p-value table we can see that the correlation between all four variables is significant on 0.05 significance level in other words the correlation between these variables is not equal to zero. From the correlation coefficient matrix, we can see all the correlations are weak. The correlation between JOBSECOK and RACE, RINCOME and CLASS are positive whereas the rest of the correlations are negative.

(Part e)

**R code :**

Using Package WriteXl

apa.cor.table(mydata)

apa.cor.table(mydata, filename="table1.doc")

#For correlation matrix

install.packages("Hmisc")

library("Hmisc")

GSS.data$RINCOME[GSS.data$RINCOME==0 | GSS.data$RINCOME > 12] <- NA

GSS.data$CLASS[GSS.data$CLASS==0 | GSS.data$CLASS > 4] <- NA

mydata <- GSS.data[c('RACE','CLASS',"RINCOME","JOBSECOK")]

mydata

mydata.rcorr = rcorr(as.matrix(mydata))

mydata.rcorr

mydata.rcorr$r

mydata.rcorr$P

**Original R output:**

*Means, standard deviations, and correlations with confidence intervals*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variable | *M* | *SD* | 1 | 2 | 3 |
|  |  |  |  |  |  |
| 1. RACE | 1.24 | 0.53 |  |  |  |
|  |  |  |  |  |  |
| 2. CLASS | 2.46 | 0.66 | -.12\*\* |  |  |
|  |  |  | [-.13, -.11] |  |  |
|  |  |  |  |  |  |
| 3. RINCOME | 9.09 | 3.40 | -.01\* | .16\*\* |  |
|  |  |  | [-.02, -.00] | [.15, .17] |  |
|  |  |  |  |  |  |
| 4. JOBSECOK | 1.67 | 0.86 | .04\*\* | -.09\*\* | -.06\*\* |
|  |  |  | [.02, .07] | [-.12, -.07] | [-.09, -.04] |
|  |  |  |  |  |  |

*Note.* *M* and *SD* are used to represent mean and standard deviation, respectively. Values in square brackets indicate the 95% confidence interval for each correlation. The confidence interval is a plausible range of population correlations that could have caused the sample correlation (Cumming, 2014). \* indicates *p* < .05. \*\* indicates *p* < .01.

RACE CLASS RINCOME JOBSECOK

RACE 1.00000000 -0.12172449 -0.01293750 0.04357999

CLASS -0.12172449 1.00000000 0.16032872 -0.09258913

RINCOME -0.01293750 0.16032872 1.00000000 -0.06458847

JOBSECOK 0.04357999 -0.09258913 -0.06458847 1.00000000

> #For p-values

> mydata.rcorr$P

RACE CLASS RINCOME JOBSECOK

RACE NA 0.000000e+00 1.811861e-02 9.520795e-04

CLASS 0.0000000000 NA 0.000000e+00 2.212008e-12

RINCOME 0.0181186076 0.000000e+00 NA 6.302493e-06

JOBSECOK 0.0009520795 2.212008e-12 6.302493e-06 NA

# **References**

General Social Survey (GSS) | NORC.org. Norc.org. Retrieved 23 August 2020, from<https://www.norc.org/Research/Projects/Pages/general-social-survey.aspx>.

Aliprantis, D., & Carroll, D. (2019). What Is Behind the Persistence of the Racial Wealth Gap?. *Economic Commentary, 2019-03.*

Brownstein, R, (2014). Are College Degrees Inherited?. *The Atlanctic, 2019-03.* <https://www.theatlantic.com/education/archive/2014/04/are-college-degrees-inherited/360532/>

# **GSS Data Analysis**

The aim of this assignment is to conduct t-test on variables of interest from the GSS data. Predictions are made regarding the variables of interest and the null hypothesis and alternative hypothesis are also stated. The results are explained after performing the test and explanation is provided as to how these align with the predictions. Furthermore, it also includes looking up how to make a correlation matrix in R and finally at the end a table is constructed containing the mean, standard deviation and correlation of four relevant variables and the results of correlation among four variables of interest are explained.

# **GSS Data**

The General Society Survey (GSS) has been studying the growing complexity of American Society and been monitoring societal change since 1972. The aim of GSS is to gather data on contemporary American society in order to monitor and explain changes and constants in behaviors, attitudes and other attributes; to examine the functioning and structure of the society as well as understand the role different relevant subgroups play. The GSS aims to gather data on contemporary American society in order to monitor and explain trends and constants in attitudes, behaviors, and attributes; to examine the structure and functioning of society in general as well as the role played by relevant subgroups and to make high-quality data easily available for students, scholars, policy makers and others. The GSS includes questions like national spending priorities, crime, marijuana use and punishment, race relations, quality of life, confidence in institutions and so on and so forth. ("General Social Survey (GSS) | NORC.org", n.d.)

1. <https://www.displayr.com/how-to-create-a-correlation-matrix-in-r/>

   <https://www.youtube.com/watch?v=dxx6azcXvRs>

   <https://stackoverflow.com/questions/23053849/creating-correlation-matrix-p-values/23056294> [↑](#footnote-ref-1)
2. <https://www.rdocumentation.org/packages/apaTables/versions/2.0.5/topics/apa.cor.table> [↑](#footnote-ref-2)
3. **How to Read a Correlation Matrix (**<https://www.statology.org/how-to-read-a-correlation-matrix/>)

   1. -1 indicates a perfectly negative linear **correlation** between two variables.
   2. 0 indicates no linear **correlation** between two variables.
   3. 1 indicates a perfectly positive linear **correlation** between two variables.

   [↑](#footnote-ref-3)