**Introduction**

This letter studies the information we have on the COVID-19 cases worldwide because the epidemic has affected billions by causing untimely deaths, and we should identify which factors currently affect COVID-19 cases worldwide.

My first research question explores if there is a relationship between the patient’s gender and whether the patient died. A recent report suggests that many death cases report “existing co-morbities” that “tend to more burdensome among men globally” (Blog | BMJ Global Health, 2020). Because of this statement, I hypothesize that there is a higher proportion of deaths among men than there are among women.

My other research question investigates whether there is a relationship between the continental region in which the patient sought care and the “Delay in days between the Onset of symptoms and seeking Hospital care” (DOH). Different regions of the world have different healthcare systems, which could affect the people’s willingness to seek medical attention for reasons such as the cost of seeking medical help. I hypothesize that there will be a significant difference in DOH based on continental regions.

My third research question asks whether there is a relationship between the patient’s age the DOH. Elderly people are more likely to seek medical attention immediately because they are more vulnerable to diseases due to weaker immune systems. I hypothesize that there will be a negative association between age and DOH.

**Methods**

The dataset comes from a published article about COVID-19 and is a case count of the dates January 13, 2020, to January 31, 2020.

For my death variable, I changed the categories from 0 and 1 to “Did Not Die” and “Died”.

For the DOH variable, I recoded values -1 and 999 to NA. I think the -1-day delay was a misinput, and the dataset codebook explained that 999 values were NA values.

For the gender variable, I combined two “male” categories and two “female” categories so that there were only two categories.

I recoded the countries into continental regions. It makes more sense to study continental regions as an explanatory variable rather than countries themselves so that I can generalize the relationships to broader regions. I used the United Nations’ M49 codes to determine regions (UN Statistic Division).

For the age variable, I recoded the 999 values into their respective NA values.

I decided to use the two-sample z-test to determine if there is a statistically significant difference in the proportion of cases resulting in death based on gender. I used the ANOVA test to determine if there is evidence of an association between continental region and DOH. I used linear regression to determine whether there is a correlation between age and delay in seeking hospital care.

**Results**

My two-sample z-test failed to reject the null hypothesis, indicating no significant difference in the proportion of death cases between males and females (Z = 1.57, p = 0.12). My ANOVA test failed to reject the null hypothesis, indicating no significant difference in DOH among continental regions (F = 2.67, p = 0.07). My linear regression test rejected the null hypothesis, indicating a negative correlation of -0.02 between age and DOH (F = 6.30, p = 0.01).

**Discussion**

My first two tests had findings contradicting my original hypotheses, implying that gender has no significant effect on whether a COVID-19 patient dies and that DOH is not significantly affected by continental regions. I could have made Type II errors, however, if the alternative hypotheses are correct. My third test’s confirmation of a negative correlation is practically insignificant because the actual strength of the correlation is extremely low (R2 = 0.02). I could have made a Type I error if there is, in reality, no correlation between the two. This study is limited because the dataset has missing values and only includes cases reported in January. Overall, none of my tests returned practical or conclusive evidence of potential factors affecting COVID-19 cases.

**Works Cited:**

2020. “Sex, gender and COVID-19: Disaggregated data and health disparities”. *Blog | BMJ Global Health*. <https://blogs.bmj.com/bmjgh/2020/03/24/sex-gender-and-covid-19-disaggregated-data-and-health-disparities/>

“Standard country or area codes for statistical use (M49)”. *UN Statistics Division*. <https://unstats.un.org/unsd/methodology/m49/#qa>

**Appendix A: Tables with descriptive statistics, tests of association, and p-values**

Table 1. Descriptive statistics for [COV Dataset]. N = 507.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Overall | Death | |
|  |  | Died | Did Not Die |
|  | N = 507 | N = 40 | N = 467 |
| Age | 47.03±17.67 | 70.25±12.73 | 44.86±16.47 |
| Gender |  | | |
| Male | 58.30% | 70% | 57.24% |
| Female | 41.70% | 30% | 42.76% |
| Continental Region |  | | |
| Asia | 92.70% | 100% | 1.93% |
| Australia | 1.78% | 0% | 92.08% |
| Europe | 3.35% | 0% | 3.64% |
| North America | 2.17% | 0% | 2.36% |

Table 2. 2-Sample z-test Results: Death by Gender. [COV Dataset]. N = 482.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Death** | **Test statistic** | **P-value** |
| Gender |  | 1.57 | 0.12 |
| Male | 9.96% |  |  |
| Female | 5.97% |  |  |

Table 3. One-Way Analysis of Variance of lag.onset.hosp by Continental Region. [COV Dataset]. N = 347

|  |  |  |  |
| --- | --- | --- | --- |
| COV Dataset | *df* | *F* | *p* |
| Continental Region | 2 | 2.67 | 0.07 |
| Residuals | 344 |  |  |
|  |  |  |  |
|  | | | |

Table 4. Linear regression results of lag.onset.hosp regressed on age in COV Dataset. N = 342.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | lag.onset.hosp (se) | 95% Confidence Interval | | p-value |
|  |  | Lower | Upper |  |
| Age | -0.02 | -0.04 | -0.01 | 0.01 |
| Constant | 3.95 | 3.02 | 4.89 | 2.41\*10-15 |
| Observations | 342 |  |  |  |
| R2 | 0.02 |  |  |  |
| F statistic | 6.30 |  |  | 0.01 |

**Appendix B:** Bi-variate plots of Gender, Continental Region and Age.

Figure 1. Association between Gender and Death. [COV Dataset]. N = 482.

A screenshot of a cell phone

Description automatically generated

Figure 2. Association between Continental Region and Lag.onset.hosp. [COV Dataset]. N = 347.

A screenshot of a cell phone

Description automatically generated

Figure 3. Association between Age and Lag.onset.hosp. [COV Dataset]. N = 342.

A screenshot of a cell phone

Description automatically generated

**Appendix C:** R code

## Elijah Chou

## 5/4/2020

## Lab #11

## Final project submission

## Title: Analysis for Potential Identification of Factors Affecting COVID-19 Cases

###Setting up Environment###

install.packages("Rmisc")

library(Rmisc)

#Set up working directory

setwd("C:/Users/ecool/Desktop/R Code")

#Import data

corona <- read.csv("C:/Users/ecool/Desktop/R Code/Data Sets/COV Dataset.csv")

####Recoding variables####

##Recoding the "death" variable

summary(corona$death) #check death variable, need to change categories from numbers to labels

corona$death2<-factor(NA, levels=c("Died", "Did Not Die")) #create new categorical variable

corona$death2[corona$death==0]<-"Did Not Die" #reassigns those who did not die to respective category

corona$death2[corona$death==1]<-"Died" #reassigns those who did die to respective category

summary(corona$death2) #verify changes to variable

##Recoding the "gender" variable

summary(corona$gender) #check gender variable, need to combine the two "male" groups into one, do same for "female" categories, and ignore 999 values

corona$gender2<-factor(NA, levels=c("male", "female")) #create duplicate "gender" variable

corona$gender2[corona$gender=="MALE" | corona$gender=="male"]<-"male" ##reassigns males to respective category

corona$gender2[corona$gender=="Female" | corona$gender=="female"]<-"female" ##reassigns females respective category

summary(corona$gender2) #verify changes to variable

##Recoding the "lag.onset.hosp" variable

summary(corona$lag.onset.hosp) #check lag.onset.hosp variable, noticed that there are two problems: a -1 value and many 999 values

corona$lag.onset.hosp2<-corona$lag.onset.hosp #create duplicate "lag.onset.hosp" variable

corona$lag.onset.hosp2[corona$lag.onset.hosp2<0]<-NA #recode all entries of this variable where delay is less than 0 to NA

corona$lag.onset.hosp2[corona$lag.onset.hosp2==999]<-NA #recode all "coded NAs" of 999 back to NA

summary(corona$lag.onset.hosp2) #verify changes to variable

##Creating the "Continental.Region" variable from "Country" variable

summary(corona$Country) #check country variable to determine which countries there are

corona$Country[corona$Country=="china"]<-"China" #recode erroneous china values

corona$Continental.Region<-factor(NA, levels=c("Australia", "Asia", "Europe", "North America")) #creates new categorical variable with the four continental regions

corona$Continental.Region[corona$Country=="Australia"]<-"Australia" #reassign values of the new Australia Continental Region

corona$Continental.Region[corona$Country=="Cambodia" | corona$Country=="China" |

corona$Country=="Hong Kong" | corona$Country=="India" |

corona$Country=="Japan" | corona$Country=="Malaysia" |

corona$Country=="Nepal" | corona$Country=="Russia" |

corona$Country=="Singapore" | corona$Country=="South Korea" |

corona$Country=="Sri Lanka" | corona$Country=="Taiwan" |

corona$Country=="Thailand" | corona$Country=="UAE" |

corona$Country=="Vietnam"]<-"Asia" #reassign values of the new Asia Continental Region

corona$Continental.Region[corona$Country=="France" | corona$Country=="Germany" |

corona$Country=="Italy" | corona$Country=="UK"]<-"Europe" #reassign values of the new Europe Continental Region

corona$Continental.Region[corona$Country=="Canada" | corona$Country=="USA"]<-"North America" #reassign values of the new North America Continental Region

summary(corona$Continental.Region) #verify successful creation of new variable

##Recoding the "age" variable

summary(corona$age) #check age variable, need to recode 999 to NA

corona$age2<-corona$age #create duplicate "age" variable

corona$age2[corona$age2==999]<-NA #recode all "coded NAs" of 999 back to NA

summary(corona$age2) #verify changes to variable

###Creating Contigency Tables###

#age vs. death

summarySE(data = corona, measurevar = "age2", na.rm = T) #overall age vs. death descriptive statistics

summarySE(data = corona, measurevar = "age2", groupvars = "death2", na.rm = T) #Provides descriptive statistics for age vs. death

summary(corona$death2)

#gender vs. death

table(corona$gender2) #contingency table of gender vs. death

201/(201+281) #proportion of women

281/(201+281) #proportion of men

prop.table(table(corona$gender2, corona$death2), margin = 2) #proportion table containing column proportions

#continental region vs. death

prop.table(table(corona$Continental.Region)) #proportion table for Continental.Region

prop.table(table(corona$Continental.Region, corona$death2), margin = 2) #proportion table of Continental Region vs. death

###Tests for Statistical Analysis###

#Table 2: Two-sample t-test for gender vs. death

Gender\_Death\_Table<-table(corona$gender2, corona$death2) #makes contingency table

addmargins(Gender\_Death\_Table) #adds margins to contingency table

prop.table(Gender\_Death\_Table, margin=1) #creates table of proportions based on row

prop.test(Gender\_Death\_Table, correct=F) #runs the 2 sample z test

sqrt(2.4566) #calculates z-score from chi-squared

#Table 3: ANOVA Test for lag.onset.hosp vs Continental.Region

anova.corona <- aov(corona$lag.onset.hosp2~corona$Continental.Region) #stores the ANOVA test results in anova.corona

summary(anova.corona) #prints summary of test results

#Table 4: Linear Regression for lag.onset.hosp vs. age

m1<-lm(corona$lag.onset.hosp2~corona$age2) #stores linear regression test in m1

summary(m1) #prints out test results

confint(m1) #provides the confidence intervals for age and constant

###Creating Plots###

#Figure 1: death vs. gender (side-by-side bar plots)

table1 <- table(corona$death2, corona$gender2) #create two-way contigency table

proptable1 <- prop.table(table1, margin=2) #create column proportions table

barplot(proptable1, beside=TRUE, legend.text = T, xlab = "Gender", ylab = "Proportion of Deaths Reported") #create bar plot using proportions with side-by-side bars

#Figure 2: lag.onset.hosp vs Continental.Region (side-by-side boxplots)

boxplot(corona$lag.onset.hosp2 ~ corona$Continental.Region, xlab = "Continental Region", ylab = "Delay in Days Between Onset of Symptoms & Seeking Hospital Care")

#Figure 3: lag.onset.hosp vs age (linear regression)

plot(corona$age2, corona$lag.onset.hosp2, xlab = "Age", ylab = "Delay in Days Between Onset of Symptoms & Seeking Hospital Care")

abline(m1) #adds the linear regression line onto scatterplot above