Sta 104: Final Project

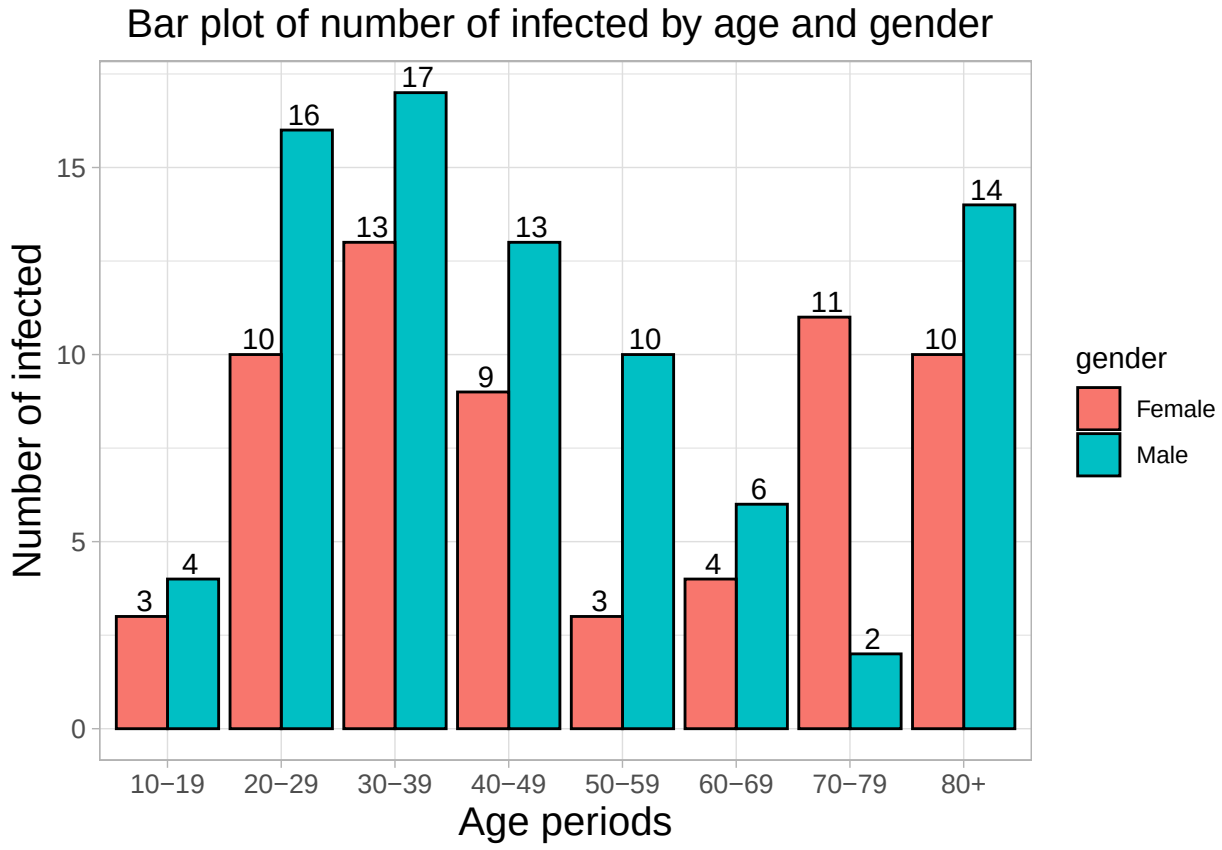
Brandon Hom, Cecilia Nguyen, Hreetu Dahal

3/11/2022

Introduction

Covid-19 is a very infectious disease that may have varying impacts depending on age and gender. Although every country is experiencing the impacts of Covid-19, we analyzed data from a rural city in Cameroon, for March 30, 2020 where data was captured on the next day to the specified date. We would like to test for independence between age and gender, specifically for the infected cases of Covid-19.

To visualize and gain a general understanding of our data, We chose to create a histogram to display the number of infected cases of each gender across age groups. From the histogram, there seems to be more cases of males infected with Covid-19 in general across all age groups. Interestingly, The category with the highest cases infected for both genders is from 30-39 years of age. Possible reasons may be that these individuals may be at the peak of their careers where they have to work with a lot of people, and also have to juggle responsibilities like finances and children. There may also be more people in the demographic, or that more people walk in the area affected.



Materials and Methods

We choose to focus on the individuals infected with Covid-19 in Cameron during those time periods. Since we're comparing two categorical variables (age and gender), we will make a contingency table. We assume that Covid-19 infected cases data are collected from observational studies where all n subjects are sampled randomly and independent, so that neither row or column totals are known beforehand.

Table 1: The Overall Number of COVID-19 Infected By Gender and Age Groups in Cameroon

	10-19	20-29	30-39	40-49	50-59	60-69	70-79	80+
Male	4	16	17	13	10	6	2	14
Female	3	10	13	9	3	4	11	10

Our hypothesis tests for independence between age and gender for infected Covid-19 cases is:

```
*$H_0$ = Age and gender are independent
vs.
*$H_A$ = Age and gender are dependent
```

To see if we require a parametric χ^2 test for independence or a non-parametric permutation test for independence, we will check if the assumptions for parametric χ^2 test for independence are fulfilled. We assume that a random sample was taken for the data, but need to compute from our contingency table if e_{ij}, the expected count if H_0 is true, is calculated by:

- $e_{ij} = \frac{n_{i.} * n_{j.}}{n}$

If e_{ij} is greater than or equal than 5 for all i,j (each cell), we will utilize a permutation based test for independence if the assumptions for parametric aren't fulfilled, where the only assumption is that a random sample of data was taken.

Results

From our table of expected counts, we observe that some cells are less than 5, which is required for the parametric χ^2 test for independence. Therefore, we will utilize a non-parametric permutation test for independence.

Table 2: Table of Expected Counts

	10-19	20-29	30-39	40-49	50-59	60-69	70-79	80+
Male	3.958621	14.70345	16.96552	12.441379	7.351724	5.655172	7.351724	13.57241
Female	3.041379	11.29655	13.03448	9.558621	5.648276	4.344828	5.648276	10.42759

Performing the parametric test, we will find our test-statistic $\chi^2_O BS$ and perform many permutations. We then obtain a p-value and compare with α to accept or reject H_0 that age and gender are independent.

Thus, the permutation based p-value is: 0.10225, which is greater than α. If in realitty gender and age for Covid-19 infected cases were independent, we would observe our data or more extreme as .11375. We then fail to reject H_0, and conclude there is evidence to support that gender and age are independent for Covid-19 infected cases.

Conclusion and Future Works

We conclude that that gender and age are independent for Covid-19 infected cases for the Cameroon area. For future works, other categorical factors may be researched to check for dependence in Covid-19 infected cases; for example, if a person is either employed or not, and their age. In addition, we may want to conduct studies on the independence of gender and age for Covid-19 death cases in the Cameroon area

{Appendix: R code used}

```
knitr::opts_chunk$set(echo = F)
library(tidyverse)
library(knitr)
age.periods <- c("10-19","20-29","30-39","40-49","50-59","60-69","70-79","80+")
male.infected <- c(4,16,17,13,10,6,2,14)
female.infected <- c(3,10,13,9,3,4,11,10)
n.infected <- c(male.infected,female.infected)
male.deaths <- c(1,2,4,8,8,3,1,9)
female.deaths <- c(1,4,11,4,2,1,9,8)
n.deaths <- c(male.deaths,female.deaths)
gender <- c(rep("Male",8),rep("Female",8))
data <- data.frame(age.periods,gender,n.infected,n.deaths)
#data
data %>%
  ggplot(aes(x=age.periods,y=n.infected,fill=gender,label=n.infected))+
  geom_col(position="dodge",col="black")+
  theme_light()+
  labs(title = "Bar plot of number of infected by age and gender ",
       y="Number of infected",
       x="Age periods")+
  theme(axis.title = element_text(size=15),
        axis.text = element_text(size=10),
        plot.title = element_text(hjust = .5, size = 15))+
  geom_text(position = position_dodge(width=.8),vjust=-.25)
infected.dat <- c(4,16,17,13,10,6,2,14,3,10,13,9,3,4,11,10)
infected_matrix <- matrix(infected.dat,nrow=2,byrow = T)
# row and column names
colnames(infected_matrix) <- age.periods
rownames(infected_matrix) <- c("Male","Female")
#convert to table for analysis
infected_table <- as.table(infected_matrix)
#infected_table
kable(infected_table, caption = "The Overall Number of COVID-19 Infected By Gender and Age Groups in Ca
# obtain expected values
the_test <- chisq.test(infected_table,correct=F)
eij <- the_test$expected
kable(eij,caption="Table of Expected Counts")
#Cleaned up dataframe for permutation test (wide to long format, shows each number of male of the age p
library(dplyr)
library(tidyr)
new_data<-data%>%
  pivot_longer(cols = c(n.infected), names_to = 'Infected') %>%
```

```
  uncount(value)  %>%
  as_tibble

bysdata<-as.data.frame(new_data) #cleaned up data
bysdata$n.deaths <- NULL
bysdata$Infected <- NULL
#bysdata
chi.sq.obs = as.numeric(the_test$statistic)
#chi.sq.obs
R = 4000
r.perms = sapply(1:R,function(i){
  perm.data = bysdata
  perm.data$gender = sample(perm.data$gender,nrow(perm.data),replace = FALSE)
  chi.sq.i = chisq.test(table(perm.data),correct = FALSE)$stat
  return(chi.sq.i)
})
perm.pval = mean(r.perms >= chi.sq.obs)
#perm.pval
```