STA 104 Exam Final Project

JONG-WOOK-CHOE, MINSOO-LEE

3/14/2022

Introduction

The COVID-19 data used here is non-publicly from a rural city in a Africa Country, Cameroon, for March 30 2020. Data were captured on the next day to the specified date. Overall statistics of infected and deceased COVID-19 patients were stratified by gender and age and expressed in table. For this project we will do analysis with **infected cases**. First we will test the relation between age and gender then comparing group difference to see which gender tends to get more infected.

Materials and Methods

Following is the contingency table of the data set,

	10-19	20-29	30-39	40-49	50-59	60-69	70-79	80+	Total
Male	4	16	17	13	10	6	2	14	82
Female	3	10	13	9	3	4	11	10	63
Total	7	26	30	22	13	10	13	24	145

	Mean
Male	10.25
Female	7.875

The data shows how many people got infected in each different age group. First looking at the data by ages from table 1, the age group "10-19" had 7 infected people but the number sharply increases in next age group "20-29". The infected number of people increases up to 30 in the age group "30-39". Then the number somewhat decreased after age group of "30-39" At the end in the age group of "80+" the number slightly increased to 24. Looking at the data by gender, the infected number of males is higher than the infected number of females.

Also, the mean of the infected male is 10.25 while the female is 7.875. The mean of male who is infected is larger than female.

Next, the following bar plot shows the data trend of the infected people by age differ by gender.

According to the histogram, the age group of "70-79" has the most gap between the number of infected people by gender. Also, the number of males tends to be larger than females in most age groups.

First, to test whether there are relation between variables. We could use parametric χ^2 test for independence when following assumption hold.

Parametric test assumption

- Random sample was taken
- $\epsilon_i \sim N(o, \sigma_\epsilon^2)$
- $e_{ij} \geq 5$ for all i,j

Following is the table of expected count on average,

$e_{ij} = \frac{n_{i\cdot} \times n_{\cdot j}}{n}$

	10-19	20-29	30-39	40-49	50-59	60-69	70-79	80+
Male	3.96	14.7	16.97	12.44	7.35	5.66	7.35	13.57
Female	3.04	11.3	13.03	9.56	5.65	4.34	5.65	10.43

However, there are some entries that $e_{ij} < 5$. Therefore assumption is violated and we have to use non parametric permutation test for independence.

Non Parametric test assumption

- Random sample is taken
- The distributions are equal (but unspecified)

As you can see they have assumptions, but they are more general.

Stating appropriate null and alternative hypothesis for independence test,

- H_0 : Variables Age, Gender are independent
- H_A : Variables Age, Gender are dependent

Next step is to do many permutations and calculate the permutation p-value

$$(\#\text{ of } \chi_i^2 \geq \chi_{OBS}^2) \,/\, R$$

Where $\chi_{OBS}^2 = 11.5636382$ and $R = 4000$

Since p-value is 0.1065, with any reasonable α we reject H_o and conclude two variables are dependent.

Now we want to check which gender is more infected by disease.

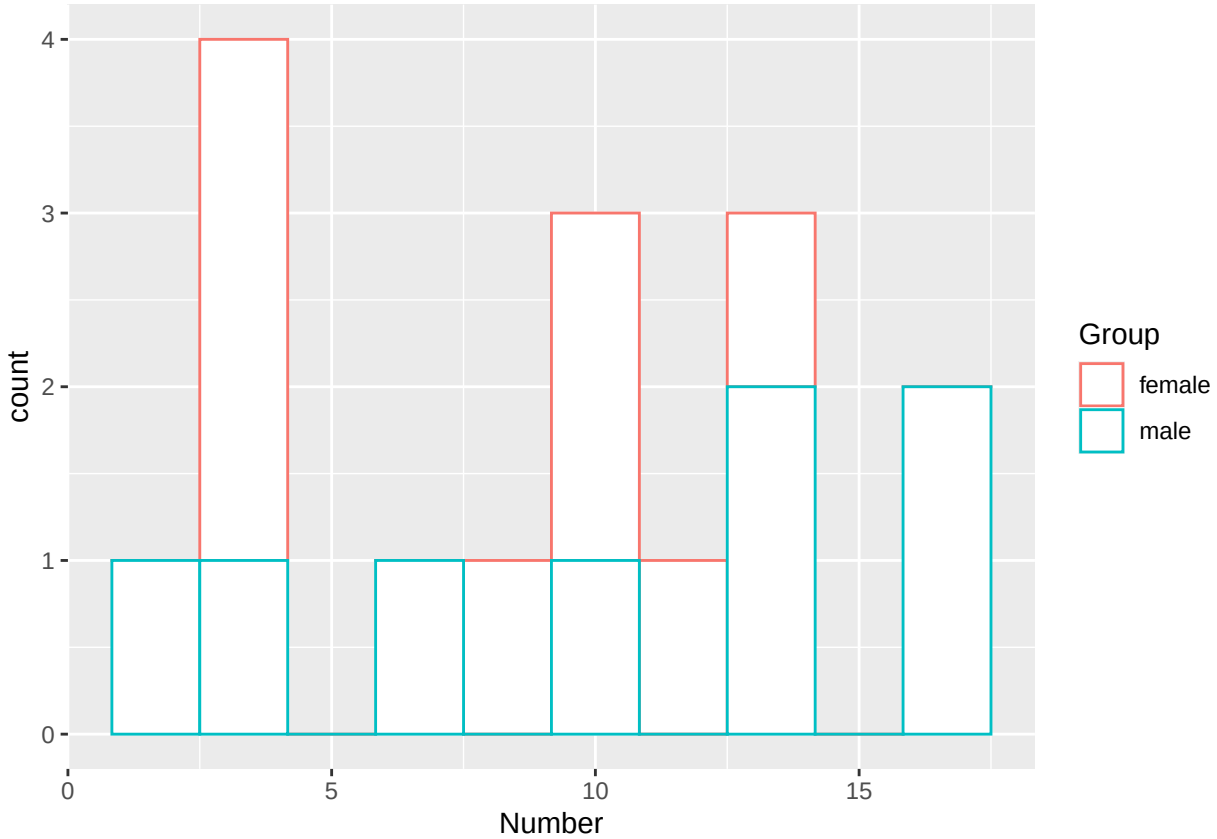
Lets call, * Group1 : female * Group2 : male

Following table is the calculated group means, group standard deviations, group average ranks of the data.

	Female	Male
Group Mean	7.875000	10.250000
Group SD	3.943802	5.675763
Rank Mean	7.250000	9.750000
Sample Size	8.000000	8.000000

Each groups sample sizes are all equal to 8. Standard deviations for each groups are 3.943802 and 5.675763 respectively. This shows that they do not have equal variance.

Now using histogram, we could see the distribution of two group,



By looking at the summary of the data and the histogram above. We cannot assume normality because the standard deviations by groups are not equal. Also outlier are present and some of the distributions are slightly skewed.

Therefore, rather than asymptotic pairwise mean difference we would use two-sample non-parametric technique.

These are few cases when we use non-parametric technique,

1) Both groups have significant outliers. → use Wilcoxon Rank Sum test or Mann–Whitney U test because they are both not effected by outliers, and generally have higher power.
2) One group has a skewed distribution, and the other has a symmetric distribution. → use Permutation test for the median. Since Median is not very effected by outliers. Also Wilcoxon Rank Sum test and Mann–Whitney works
3) Both groups have a symmetric distribution → use Permutation test for the mean. It tends to have higher power when the distributions are symmetric
4) There is not a clear indication to what statistic to use in order to compare the distributions. → use Kolmogorov–Smirnov test. This is used when we want to directly compare distributions through their values, rather than comparing them through a statistic.

Also, there is Kruskal–Wallis test **however**, we can't use it because of anova assumption because we are comparing only two groups.

Wilcoxon Rank Sum test tends to have higher power when the distribution is skewed or outliers are present. Since assigning ranks essentially removes all influence of both issues it outperform Permutation tests. Therefore we will use Wilcoxon Rank Sum test to compare pairwise mean difference.

Wilcoxon Rank Sum test assumption

- A random sample was taken from each group.
- Groups are independent.

Stating appropriate null and alternative hypothesis for wrs test,

- $H_0 : \mu_1 - \mu_2 \leq 0$
- $H_A : \mu_1 - \mu_2 > 0$

Our test statistic requires the following steps

1) Combine the $m + n$ values into one group
2) Calculate the rank for each data point $R(x_i) = \#$ of data $\leq x_i$ $i = 1, \cdots, m + n$. If there are ties, average the ranks of the tied observations, and assign the tied values that are rank.
3) Calculate the total rank in group 1. This is our test-statistic, $W_{OBS} = \sum_{\text{group 1}} R(x_i)$

To find the exact p=value we calculate all $(\frac{m+n}{n})$ permutations of the two groups, and calculate the distribution of W_i =sum of ranks in group 1. Then

$$(\#\text{ of } W_i \geq W_{OBS})/(\frac{m+n}{n})$$

Since p-value is 0.1553 at any reasonable value of α, we would also fail to reject null, and cannot conclude that female has larger mean than male.

Since we fail to reject null lets test other way around whether male has larger mean than female.

Since p-value is 0.8562 at any reasonable value of α, we would fail to reject null and cannot conclude that male has larger mean than female.

Since p-value is 0.3106 at any reasonable value of α, we would fail to reject null and cannot conclude that mean of male and female have difference.

Results

Relation between Two Groups

- Interpreting P-value: If two variable **gender** and **age** are independent to each other, we would observe our data or more extreme with probability 0.1065.
- Since P-value (0.1065) $<$ any of our available α, we can reject H_o that the two variable **gender** and **age** are independent. Therefore, we conclude that two variable gender and age are **dependent**.

Distribution between gender, Male and Female

- Interpreting P-value ($H_0 : \mu_1 - \mu_2 \leq 0$, where 1:Female and 2:Male): If male average infected number is larger or equal to female average infected number, we would observe our data or more extreme with probability 0.1553.
- Since p-value:0.1553 $>$ any of our available α, we can not reject H_o that male average infected number is larger or equal to female average infected number. Therefore, we do not have significant evidence to conclude that average number of infected female are larger than average number of infected male.
- Interpreting P-value ($H_0 : \mu_1 - \mu_2 \geq 0$, where 1:Female and 2:Male): If average number of infected female is larger or equal to average number of infected male, we would observe our data or more extreme with probability 0.8562.
- Since p-value:0.8562 $>$ any of our available α, we can not reject H_o that female average infected number is larger or equal to male average infected number. Therefore, we do not have significant evidence to conclude that average number of infected male are larger than average number of infected female.
- Interpreting P-value ($H_0 : \mu_2 - \mu_1 = 0$, where 1:Female and 2:Male): If average number of infected female is larger or equal to average number of infected male, we would observe our data or more extreme with probability 0.3106.
- Since p-value:0.3106 $>$ any of our available α, we can not reject H_o that female average infected number is equal to male average infected number. Therefore, we do not have significant evidence to conclude that average number of infected male and female is different.

Conclusion and Future work

Throughout our analysis, we had proved that there is a relationship between variables age and gender in an overall number of COVID-19 infected cases in Cameroon. This indicates that certain age and gender groups had been more infected by the disease.

Additionally, we investigated the pairwise differences in gender groups. Overall statistics table of infected patients stratified by gender and age tend to be a higher number in male than female. We wanted to test whether the number of males tends to be larger than females in most age groups. Using the Wilcoxon rank-sum test we calculated the p-value, however, because of the small sample size it is hard to reach any conclusions but one thing clear is that there was no significant difference between gender getting infected by disease.

Furthermore, we could also discuss the overall number of deceased cases by gender and age groups. Using the same method above, we could figure out whether there's any significant difference from infected cases.

Appendix

```
library(plyr)
library(ggplot2)
library(survival)
library(coin)
```

```
raw1 = matrix(data = c(4,3,16,10,17,13,13,9,10,3,6,4,2,11,14,10),nrow = 2)
raw2 = matrix(data = c(1,1,2,4,4,11,8,4,8,2,3,1,1,9,9,8),nrow = 2)
rownames(raw1) = c("Male","Female")
rownames(raw2) = c("Male","Female")
colnames(raw1) = c("10-19", "20-29", "30-39", "40-49", "50-59", "60-69", "70-79", "80+")
colnames(raw2) = c("10-19", "20-29", "30-39", "40-49", "50-59", "60-69", "70-79", "80+")

longify_xtab <- function(x) {
  nm <- names(x)
  # Convert to table
  x_tab <- as.table(as.matrix(x))
  # Just in case there are now rownames, required for conversion
  rownames(x_tab) <- nm
  # Use appropriate method to get a df
  x_df <- as.data.frame(x_tab)

  # Restructure df in a painful and unsightly way
  data.frame(lapply(x_df[seq_len(ncol(x_df) - 1)], function(col) {
    rep(col, x_df$Freq)
  }))
}

df_i = longify_xtab(raw1)
df_i$Var1=ifelse(df_i$Var1=="A","male","female")
df_i=rename(df_i,c('Var1'= 'Gender', 'Var2'= 'Age'))
df_d = longify_xtab(raw2)
df_d$Var1=ifelse(df_d$Var1=="A","male","female")
df_d=rename(df_d,c('Var1'= 'Gender', 'Var2'= 'Age'))
ggplot(df_i, aes(x=Age,fill=Gender))+geom_bar(position = 'dodge')
n=sum(raw1)
ni. = rowSums(raw1)
n.j = colSums(raw1)

the.test = chisq.test(raw1,correct = FALSE,simulate.p.value=TRUE)
eij = round(the.test$expected,digits=2)
chi.sq.obs = as.numeric(the.test$statistic)
R = 4000
r.perms = sapply(1:R,function(i){
  perm.data = df_i
  perm.data$Age = sample(perm.data$Age,nrow(perm.data),replace = FALSE)
  chi.sq.i = chisq.test(table(perm.data),correct = FALSE,simulate.p.value=TRUE)$stat
  return(chi.sq.i)
})
perm.pval = mean(r.perms >= chi.sq.obs)
some.numbers = c(4,16,17,13,10,6,2,14,3,10,13,9,3,4,11,10)
some.groups = c("male","male","male","male","male","male","male","male","female","female","female","fema
some.data = data.frame(Number = some.numbers,Group = some.groups)

some.data$Rank = rank(some.data$Number, ties = "average")

Group.order = aggregate(Number ~ Group, some.data, mean)$Group
Xi = aggregate(Number ~ Group, some.data, mean)$Number
si = aggregate(Number ~ Group, some.data, sd)$Number
```

```
Ri = aggregate(Rank ~ Group, some.data, mean)$Rank
ni = aggregate(Number ~ Group, some.data, length)$Number
results = rbind(Xi,si,Ri,ni)
rownames(results) = c("Group Mean","Group SD","Rank Mean","Sample Size")
colnames(results) = as.character(Group.order)
ggplot(some.data, aes(x=Number,color=Group))+geom_histogram(fill='white',bins=10)
wrstestl=wilcox_test(Number ~ as.factor(Group),some.data, distribution = "exact",alternative = "less")
wrstestg=wilcox_test(Number ~ as.factor(Group),some.data, distribution = "exact",alternative = "greater
wrstestt=wilcox_test(Number ~ as.factor(Group),some.data, distribution = "exact",alternative = "two.sid
```