STA 104 Midterm 2 Project Report

Radhika Kulkarni, Kevin Xu

2/26/2021

Class: STA 104 - Nonparametric Statistics
Instructor: Amy T. Kim
Email: rrkulkarni@ucdavis.edu
kivxu@ucdavis.edu
Student IDs: 917212169, 913288760

Question 1: K or more groups

I. Introduction.

The following paper addresses the question of whether State has an effect on the number of deaths due to COVID-19. Each state has its own methods and infrastructure for COVID testing, policy, and treatment, so we expect there may be some variability in the number of COVID-19 deaths due to State. We chose the states with the largest populations because their COVID-19 policies affect a large number of people and they would be forced into a similar situation in terms of COVID 19 deaths and thus are comparable. These states are California, Texas, New York, and Florida. A claim we would like to test is that the states have a significant effect on the number of deaths due to COVID 19. We are interested in the result of testing this claim because we would be able to better understand or at least solidify a starting point for understanding which method of dealing with COVID 19's effects would result in lower number of deaths. In large population states this will be even more impactful if we find some difference because then public policy can be tailored to more effectively reduce the number of deaths due to COVID-19. We will use the statistical technique of Kruskal Wallis to determine which state has a more effective technique for lower number of COVID 19 deaths.

II. Summary of Data

This paper utilizes COVID-19 death count data of the four states with the largest population in the US: California, Texas, New York, and Florida. These data are provisional from the CDC database.

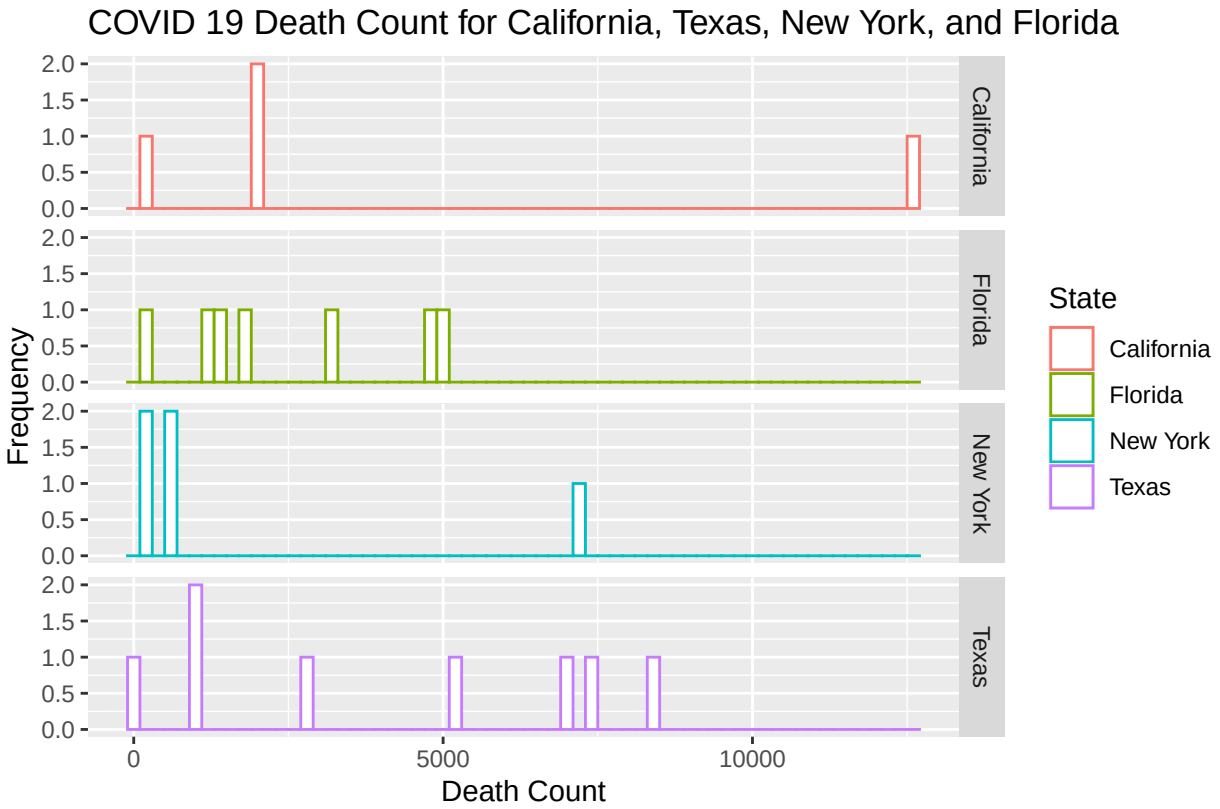

Figure 1: Histograms for three out of four states are skewed right with outliers and heavy tails.

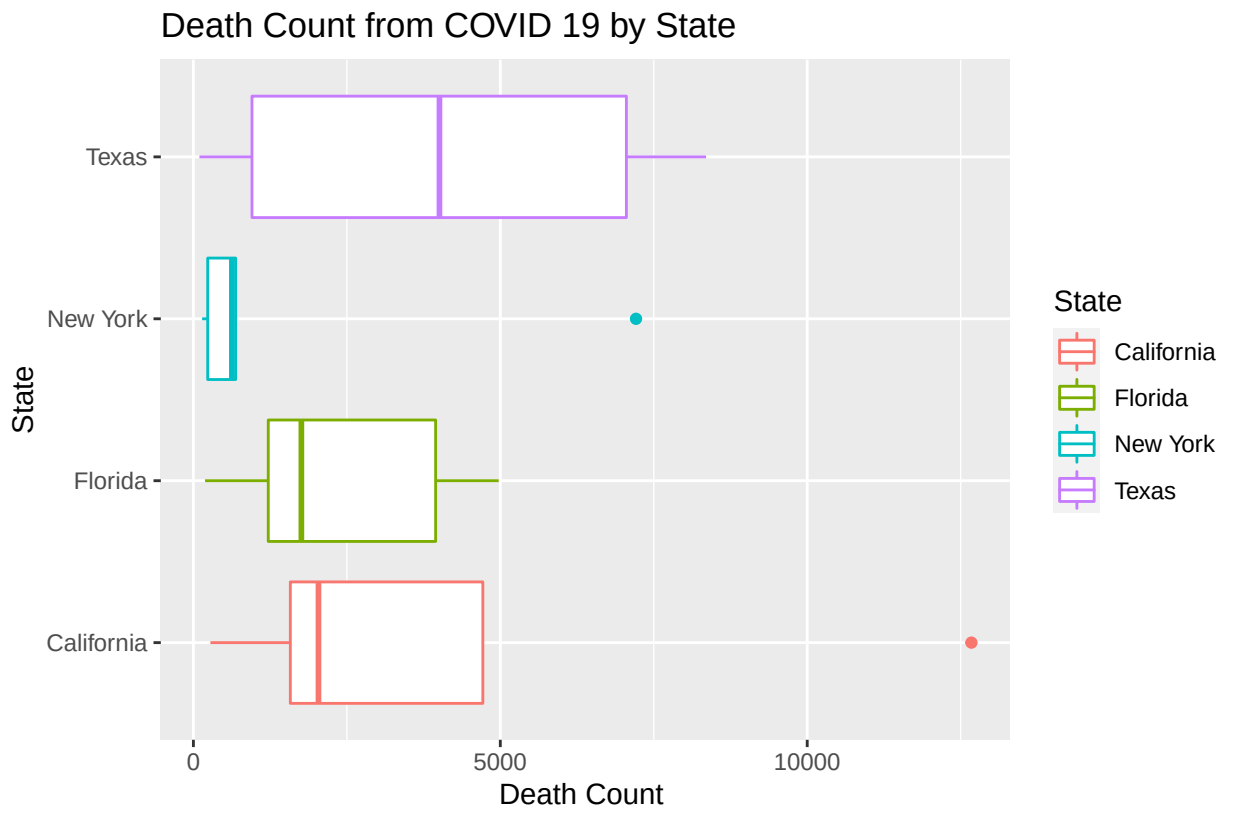

Figure 2: Two of the four states display clear outliers, and three of the four show severe skewness.

	California	Florida	New York	Texas
Group Mean	4256.750	2465.00	1780.200	4083.250
Group SD	5673.232	1863.53750	3045.238	3282.648
Rank Mean	14.00	12.42857	8.00	14.625
Sample Size	4.00	7.00	5.00	8.00

Table 1: The sample sizes for each of the states are small, which is one of the reasons why a nonparametric test is more appropriate for this dataset.

The group mean deaths for California (sample size = 4) is 4256.75, with a rank mean of 14.00 and standard deviation of 5673.232. The group mean deaths for Florida (sample size = 7) is 2465.00, with a rank mean of 12.42 and standard deviation of 1863.538. The group mean deaths for New York (sample size = 5) is 1780.20, with a rank mean of 8.00 and standard deviation of 3045.238. The group mean deaths for Texas (sample size = 8) is 4083.250, with a rank mean of 14.63 and standard deviation of 3282.648 deaths. New York has a lower mean (1780.20) than the other states, which may indicate a better COVID strategy and possibly a lower number of deaths as well. However, the variance is rather high (3045.238) which in a small sample could mean that any inferences we make are not very reliable. California, on the other hand, has the largest mean (4256.750), with Texas not far behind (4083.250), and these states also have a rather high variance which may lead us to think that there is again an uncertainty in how many deaths there actually are due to the small sample sizes. The high mean deaths of both states could indicate that they could be the result of a less effective COVID response than that of New York's.

First we check the assumptions of ANOVA to see if we can use a parametric test. We use the Shapiro-Wilkes Test and Levene Test to test for non-normality and constant variance, respectively. The p-value for the

Shapiro-Wilkes test, 0.001395, is less than common values of α, which is strong evidence for non-normality. Although the Levene Test shows that we have constant variance (p-value = 0.09158 > α = 0.05), the assumption for normality for the parametric test is violated. Thus we must use the nonparametric test of permutation or Kruskal-Wallis.

We use the nonparametric technique of Kruskal Wallis because as we can see from our data visualizations, three of the four states : California, New York, and Florida have visible skewness to the right, New York and California have clear outliers, and California, Texas, and Florida appear to have heavy tails. Using ranks tends to make inferences that are more robust to outliers, so a rank-based method such as Kruskal-Wallis Test allows for a result that has more power than a permutation test. Because the sample sizes of 4,5,7,8 are not large enough, we cannot use the large sample approximation to the Kruskal Wallis Test.

III. Analysis.

Our null hypothesis is that the distributions of the death counts for each of the four states are the same, and our alternative hypothesis is that there is at least one difference between the distributions of the death counts for each state. Essentially we are testing whether our claim that State COVID-19 response strategies have an effect on the number of COVID 19 deaths (the death count differs significantly between the four states) is valid, or else State COVID-19 response strategies have no effect on death count. We used the largest population states of California, New York, and Florida, and Texas to measure the differences in State.

We conducted the Kruskal-Wallis Test and obtained a test statistic of 2.928214, and an exact p-value of 0.4263333.

IV. Interpretation.

If the test statistic were 2.928214 or greater, assuming that there are no differences between state COVID 19 response effect on the death count, we would observe our data or more extreme with probability 0.4263333.

Since we failed to reject the null hypothesis, we cannot conduct simultaneous inference for pairwise differences between the states.

Because our p-value = 0.4263333 > α = 0.05, we fail to reject the null hypothesis. We conclude that the death count does not differ significantly between the four states.

V. Conclusion.

The goal of this paper was to determine whether State COVID-19 response strategies have an effect on the number of COVID 19 deaths for the four states with the largest populations in the United States, which are California, Texas, New York, and Florida. From the death count of each state, we determined using Kruskal Wallis test that the distributions of the death counts for each of the four states are the same, and in fact the claim is false. As a result, we cannot say that the COVID-19 protocols and regulations of the states with the largest populations have an effect on the death count. It seems like that it may require more testing and data in the future to determine whether state COVID-19 response strategies are effective.

Question 2: Tests for Independence

I. Introduction.

The following paper addresses whether Age (in years) and Sex (Female, Male) are independent variables regarding Deaths involving COVID-19. COVID-19 is a contagious airborne virus that instigated a global pandemic. It has afflicted people throughout the globe starting from September 2019 until the present day.

Data corresponding to reported COVID-19 related deaths of people aged 18 years and older, within the time period March 2020 to February 2021, was utilized for analysis to investigate whether Age and Sex are independent variables regarding death by a COVID-19 related ailment. Adults (18 years and older) were selected for analysis to study the risk of death from a Covid-19 related ailment because of confidentiality for people under the age of 18 years. Additionally, analysis of the independence of the variables Age and Sex regarding death as a result of COVID-19 has broader implications on whether age matters for how at risk a person is of dying from COVID-19, for a given sex, and whether sex matters for how at risk a person is of dying from COVID-19 for a given age.

This paper uses the permutation test for independence between Age group and Sex regarding Deaths involving COVID-19. We will find a permutation p-value, and if sufficient evidence is given such that we can reject the null hypothesis (that groups Age and Sex are independent), we will subsequently identify what the dependency is: using multiple comparisons in contingency tables and the Tukey's HSD inspired cutoff values.

II. Summary of Data

The paper utilizes data of a sample of 478678 COVID-19 related deaths, of subjects 18 years or older. The cleaned dataset provided by STA 104, `CovidB` originated from the provisional COVID-19 Death Counts by Sex, Age, and State from the CDC. It is assumed that each month is independent.

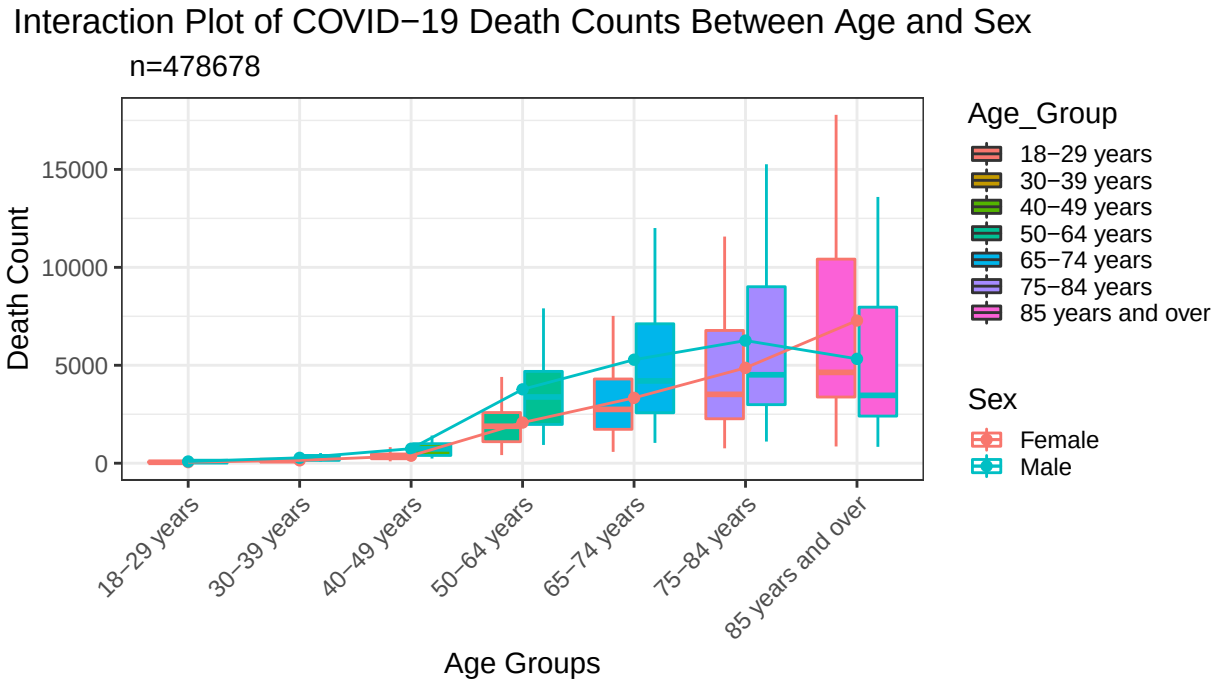

Figure 1: Variability of Death counts appears to be much larger for elder Age−groups.
In any given Age−group other than the age groups of 85 and older,
Males have a higher death count than Females.
Number of deaths per age group do not appear to be the same regardless of sex, and number of deaths per sex do not appear to be the same regardless of age.
Death counts are skewed right for Age−groups over the age of 65.

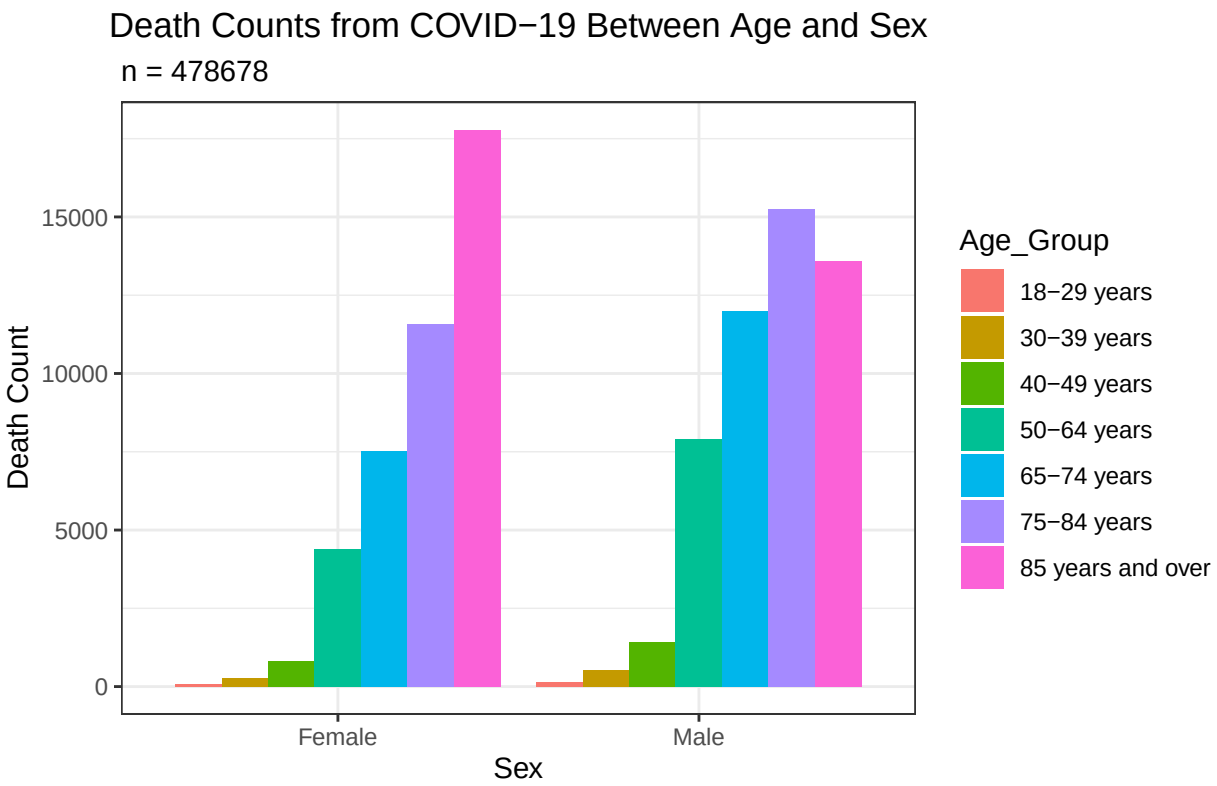

Figure 2: Number of deaths per age group do not appear to be the same regardless of sex and the number of deaths per sex do not appear to be the same regardless of age.

	Sex	
Age Group	Female	Male
18-29 years	765.9052	918.0948
30-39 years	2286.8000	2741.2000
40-49 years	6131.7895	7350.2105
50-64 years	31906.0449	38245.9551
65-74 years	47045.4068	56393.5932
75-84 years	60741.1921	72810.8079
85 years and over	68831.8614	82509.1386

Table 2: All e_{ij} are greater than or equal to five, which could possibly indicate that we may use the parametric test for independence.

We determine if a parametric χ^2 independence test is appropriate. The assumptions are that samples are randomly selected, the expected count on average (assuming Age Group and Sex are independent) is greater than or equal to five, and that the magnitude of all n_{ij} is not very different. It is assumed from the dataset `CovidB` that all months are independent, the expected counts on average are all greater than 5 (as shown in Table 2), but n_{ij} have very different magnitudes (differing by up to 86762 deaths as seen in Table 3). This indicates that our test statistic: $\chi^2_s = \sum_{i,j} \frac{(n_{ij}-e_{ij})^2}{e_{ij}}$, may not have a χ^2 distribution. Hence, the parametric χ^2 independence test is not appropriate and we use the nonparametric technique of permutation test for independence for the data at hand.

Contingency Table

	Sex		
Age Group	Female	Male	
18-29 years	640	1044	1684
30-39 years	1736	3292	5028
40-49 years	4582	8900	13482
50-64 years	24893	45259	70152
65-74 years	40030	63409	103439
75-84 years	58426	75126	133552
85 years and over	87402	63939	151341
	217709	260969	478678

Table 3: Females of younger age groups have a smaller number of deaths that keeps increasing with age group until age of 85 years and over, where Female deaths overtake Male deaths.

	Female	Male
(18-29)	0.3800475	0.6199525
(30-39)	0.3452665	0.6547335
(40-49)	0.3398606	0.6601394
(50-64)	0.3548438	0.6451562
(65-74)	0.3869914	0.6130086
(75-84)	0.4374775	0.5625225
(85+)	0.577517	0.422483

Table 4: Probability of Death from COVID-19 for by Age group given a subject is a Male or a Female

	18-29	30-39	40-49	50-64	65-74	75-84	85+
Female	0.002939704	0.007973947	0.021046443	0.114340702	0.183869293	0.268367408	0.401462503
Male	0.004000475	0.012614525	0.034103667	0.173426729	0.242975219	0.287873272	0.245006112

Table 5: Probability of Death from COVID-19 for by Sex given a subject is in a particular Age group

III. Analysis.

Our null hypothesis is that the variables Age and Sex are independent variables regarding Deaths involving COVID-19, and our alternative is that the variables Age and Sex are dependent regarding Deaths involving COVID-19. We conducted a nonparametric permutation test for independence and obtained a test statistic of $\chi_s^2 = 15097.52$. We obtained a permutation p-value of ~0 if rows (Sex) were fixed, and a permutation p-value of ~0 if columns (Age Group) were fixed.

When the permutation distribution was calculated by fixing the row totals (COVID-Deaths by Sex) and shuffling them into all columns (Age-groups), there were $\frac{478678!}{217709!*260969!}$ ways to shuffle columns and fix rows, and $\frac{478678!}{1684!*5028!*13482!*70152!*103439!*133552!*151341!}$ ways to shuffle rows and fix columns. Total COVID-19 deaths per sex were shuffled into Age-groups R = 3000 times in the permutation test for independence.

Further, multiple comparisons in contingency tables were constructed to determine whether there exists significant pairwise differences between all $7 * (7-1)/2 = 21$ combinations of people of different age groups of 18-29, 30-39, 40-49, 50-64, 65-74, 75-84, and 85 years and older given that they are of the same sex. Additionally we investigate pairwise differences between people of different sex given they are in the same age group.

We find significant pairwise differences based on a Tukey's HSD inspired cutoff value at a 95% confidence level(R=2000) of $q_1^*(0.05) = 2.866902$ for comparisons between age groups, and $q_2^*(0.05) = 2.626315$ for comparisons between sexes (Female vs Male).

	Female	Male
18-29 vs 30-39	2.480822	-2.480822
18-29 vs 40-49	3.122544	-3.122544
18-29 vs 50-64	2.05256	-2.05256
18-29 vs 65-74	-0.56764	0.56764
18-29 vs 75 -84	-4.70327	4.70327
18-29 vs 85+	-16.1837	16.1837
30-39 vs 40-49	0.656984	-0.656984
30-39 vs 50-64	-1.3174	1.3174
30-39 vs 65-74	-5.80225	5.80225
30-39 vs 75-84	-12.8904	12.8904
30-39 vs 85+	-325363	325363
40-49 vs 50-64	-3.19979	3.19979
40-49 vs 65-74	-10.3369	10.3369
40-49 vs 75-84	-21.6935	21.6935
40-49 vs 85+	-53.1015	53.10152
50-64 vs 65-74	-13.1995	13.1995
50-64 vs 75-84	-35.5888	35.5888
50-64 vs 85+	-97.9033	97.9033
65-74 vs 75-84	-24.4785	24.4785
65-74 vs 85+	-94.8423	94.8423
75-84 vs 85+	-74.9073	74.90725

Table 6: There are 16 significant pairwise differences, and 5 insignificant ones, based on cutoff value 2.866902.

	18-29	30-39	40-49	50-64	65-74	75-84	85+
Female vs Male	-6.172323	-15.68195	-27.19006	-57.55848	-49.47539	-14.98323	115.9236

Table 7: All seven pairwise differences are significant based on the cutoff value of 2.626315.

IV. Interpretation.

In the nonparametric χ^2 permutation test, if the test statistic were 15097.52 or greater, assuming age and sex were independent variables regarding deaths involving COVID-19, we would observe our data or more extreme with probability ~0. We believe it is possible this extreme p-value was caused by time-dependency, due to our extensive time period of 12 months (March 2020- February 2021).

Because our p-value = $\sim 0 < \alpha = 0.05$ for fixed rows (Age) with randomly shuffled columns and for fixed columns (Sex) with shuffled rows, we reject the null hypothesis. We conclude that Age and Sex are dependent variables regarding deaths involving COVID-19.

We then identified the dependency using multiple comparisons in contingency tables and the associated Tukey's HSD inspired cutoff values. At the cutoff values, there exists 16 significantly different pairwise differences out of 21 when different age groups are compared for given sex, and there exists 7 significantly different pairwise differences out of 7 when sexes are compared for given ages. This is shown in Tables 6 and 7.

As supported by the conditional probabilities of Tables 4 and 5, this suggests that death proportions of Age groups: 18-29 years, 30-39 years, 40-49 years, 50-64 years, 65-74 years, 75-84 years, and 85 years and older, differ significantly between Females and Males with more males dying from COVID-19 related illness for age groups 18-29 years to 75-84 years, and with more females dying from COVID-19 related illness for age group 85 years and older.

Additionally, death proportions of Sex groups: Female and Male, differ significantly between the age groups: (18-29 years and 40-49 years) with more 18-29 year olds than 40-49 year olds dying from COVID-19 related

illness for Females, and more 40-49 year olds dying for Males.

(18-29 years and 75-84 years) with more 75-84 year olds than 18-29 year olds dying from COVID-19 related illness for Females, and more 18-29 year olds for Males.

(18-29 years and 85 years and older) with more 85 year olds and older than 18-29 year olds dying from COVID-19 related illness for Females, and more 18-29 year olds for Males.

(30-39 years and 65-74 years) with more 65-74 year olds than 30-39 year olds dying from COVID-19 related illness for Females, and more 30-39 year olds for Males.

(30-39 years and 75-84 years) with more 75-84 year olds than 30-39 year olds dying from COVID-19 related illness for Females, and more 30-39 year olds for Males.

(30-39 years and 85 years and older) with more 85 year olds and older than 30-39 year olds dying from COVID-19 related illness for Females, and more 30-39 year olds for Males.

(40-49 years and 50-64 years) with more 50-64 year olds than 40-49 year olds dying from COVID-19 related illness for Females, and more 40-49 year olds for Males.

(40-49 years and 65-74 years) with more 65-74 year olds than 40-49 year olds dying from COVID-19 related illness for Females, and more 40-49 year olds for Males.

(40-49 years and 75-84 years) with more 75-84 year olds than 40-49 year olds dying from COVID-19 related illness for Females, and more 40-49 year olds for Males.

(40-49 years and 85 years and older) with more 85 year olds and older than 40-49 year olds dying from COVID-19 related illness for Females, and more 40-49 year olds for Males.

(50-64 years and 65-74 years) with more 65-74 year olds than 50-64 year olds dying from COVID-19 related illness for Females, and more 50-64 year olds for Males.

(50-64 years and 75-84 years) with more 75-84 year olds than 50-64 year olds dying from COVID-19 related illness for Females, and more 50-64 year olds for Males.

(50-64 years and 85 years and older) with more 85 year olds and older than 50-64 year olds dying from COVID-19 related illness for Females, and more 50-64 year olds for Males.

(65-74 years and 75-84 years) with more 75-84 year olds than 65-74 year olds dying from COVID-19 related illness for Females, and more 65-74 year olds for Males.

(65-74 years and 85 years and older) with more 85 year olds and older than 65-74 year olds dying from COVID-19 related illness for Females, and more 65-74 year olds for Males.

(75-84 years and 85 years and older) with more 85 years and older than 75-84 year olds dying from COVID-19 related illness for Females, with more 75-84 year olds for Males.

V. Conclusion.

The goal of this paper was to determine if Age and Sex are independent variables regarding COVID-19 related deaths. The timeframe selected for analysis is March 2020 to the end of February 2021, and Age-groups selected for analysis are those 18 years and older who have died from COVID-19. We determined using the nonparametric χ^2 permutation test that Age and Sex are independent variables regarding death by a COVID-19 related ailment. We rejected the null hypothesis that variables Age and Sex regarding COVID-19 related deaths are independent, and we conclude that the variables Age and Sex are dependent regarding the chosen Age group COVID-19 related deaths.

It appears that people who are the same age are not equally at risk of dying by an ailment related to Covid-19 given that they are of different sexes, and people of the same sex are not equally at risk of dying from an ailment related to Covid-19 given they belong to different age groups.

Next, we tested for level of dependency and found that the proportion of males and females who die from COVID-19 related illness in all age group (18-85+) is not the same, with males tending to die more often than females for age groups 18-84 years, and females tending to die more often than males for the age group 85 years and older. Out of the majority of the pairwise differences between age groups, we found that the pairwise differences were significant. This has far reaching implications on the impact of age and sex on risk of death for adults. Inferences extracted from this analysis could inform at-risk age-sex populations of the impending risks they face and potentially save lives.

Appendix Code

```
knitr::opts_chunk$set(echo = TRUE)
library(kableExtra)
CovidA <- read.csv("C:/Users/kulra/Desktop/CovidA.csv")
stateCovidA <- CovidA$State
deathCovidA <- CovidA$Death

# subsetting
# four
# states
# with
# largest
# population
# from
# main
# dataset
# CovidA
ca.data <- CovidA[CovidA[, 3] == "California", ]
ca.deaths <- ca.data$Death
ny.data <- CovidA[CovidA[, 3] == "New York", ]
ny.deaths <- ny.data$Death
fl.data <- CovidA[CovidA[, 3] == "Florida", ]
fl.deaths <- fl.data$Death
tx.data <- CovidA[CovidA[, 3] == "Texas", ]
tx.deaths <- tx.data$Death


# new
# dataframe
# with
# our
# chosen
# four
# states
fourStates <- rbind.data.frame(ca.data, tx.data, ny.data, fl.data)
fourStates$Rank = rank(fourStates$Death, ties = "average")  #rank column
# View(fourStates)
State <- fourStates$State
State <- as.factor(State)
Death <- fourStates$Death

library(ggplot2)

plot <- ggplot(fourStates, aes(x = Death)) + geom_histogram(binwidth = 200,
    aes(color = State), fill = "white") + facet_grid(State ~
    .) + ggtitle("COVID 19 Death Count for California, Texas, New York, and Florida") +
    xlab("Death Count") + ylab("Frequency")

plot <- plot + labs(caption = "Figure 1: Histograms for three out of four states are skewed right with
    theme(plot.caption = element_text(hjust = 0.5))
plot
```

```r
library(viridis)
library(ggplot2)
plot2 <- ggplot(fourStates, aes(y = Death, x = State)) + geom_boxplot(aes(color = State)) +
    ylab("Death Count ") + xlab("State") + ggtitle("Death Count from COVID 19 by State") +
    coord_flip() + labs(caption = "Figure 2: Two of the four states display clear outliers, and three o
    theme(legend.position = "right")

plot2 <- plot2 + scale_fill_grey(start = 0.2, end = 0.8) + theme(plot.caption = element_text(hjust = 0.5
plot2


Group.order = aggregate(Death ~ State, data = fourStates, mean)$State
Xi = aggregate(Death ~ State, data = fourStates, mean)$Death
si = aggregate(Death ~ State, data = fourStates, sd)$Death
Ri = aggregate(Rank ~ State, data = fourStates, mean)$Rank
ni = aggregate(Death ~ State, data = fourStates, length)$Death
results = rbind(Xi, si, Ri, ni)
rownames(results) = c("Group Mean", "Group SD", "Rank Mean",
    "Sample Size")
colnames(results) = as.character(Group.order)
results

shapiro.test(fourStates$Death)
library(car)
# Levene's
# test
# with
# mean
# as
# the
# center
# for
# each
# group
leveneTest(fourStates$Death ~ as.factor(fourStates$State), center = mean)

SR.2 = var(fourStates$Rank)
N = nrow(fourStates)
K = length(unique(fourStates$State))

KW.OBS = 1/SR.2 * sum(ni * (Ri - (N + 1)/2)^2)  #Note, this assumes you calculate ni and Ri above
set.seed(3000)
R = 3000
many.perms.KW = sapply(1:R, function(i) {
    permuted.data = fourStates  #So we don't overwrite the original data
    permuted.data$State = sample(permuted.data$State, nrow(permuted.data),
        replace = FALSE)  #Permuting the groups
    SR.2 = var(permuted.data$Rank)
    ni = aggregate(Rank ~ State, data = permuted.data, length)$Rank
    Ri = aggregate(Rank ~ State, data = permuted.data, mean)$Rank
    KW.i = 1/SR.2 * sum(ni * (Ri - (N + 1)/2)^2)
    return(KW.i)
})
```

```r
p.value = mean(many.perms.KW > KW.OBS)
p.value

# ------------------------------------------------------------------------------
# pairwise
# differences-
# we
# do
# not
# test
# for
# them
# because
# when
# we
# fail
# to
# reject
# the
# null
# hypothesis
# we
# retain
# that
# there
# are
# no
# pairwise
# differences
all.diff = as.numeric(dist(Ri, method = "manhattan"))
# There
# are
# K(K-1)
# / 2
# pairwise
# differences,
# or
# 4*(4-1)
# / 2
# = 6
# total.
names(all.diff) = c("California vs Florida", "California vs New York",
    "California vs Texas", "Florida vs New York", "Florida vs Texas",
    "New York vs Texas")
all.diff
K = length(unique(fourStates$State))
alpha = 0.05
g = K * (K - 1)/2
BON12 = qnorm(1 - alpha/(2 * g)) * sqrt(SR.2 * (1/ni[1] + 1/ni[2]))
BON13 = qnorm(1 - alpha/(2 * g)) * sqrt(SR.2 * (1/ni[1] + 1/ni[3]))
BON23 = qnorm(1 - alpha/(2 * g)) * sqrt(SR.2 * (1/ni[2] + 1/ni[3]))
all.BON = c(BON12, BON13, BON23)
```

```r
HSD12 = qtukey(1 - alpha, K, N - K) * sqrt((SR.2/2) * (1/ni[1] +
    1/ni[2]))
HSD13 = qtukey(1 - alpha, K, N - K) * sqrt((SR.2/2) * (1/ni[1] +
    1/ni[3]))
HSD23 = qtukey(1 - alpha, K, N - K) * sqrt((SR.2/2) * (1/ni[2] +
    1/ni[3]))
all.HSD = c(HSD12, HSD13, HSD23)

all.crits = rbind(all.diff, all.BON, all.HSD)
all.crits
# ---------------------------------------------------------------------
CovidB <- read.csv("C:/Users/kulra/Desktop/CovidB.csv")

# Subsetting
# for
# Age
# Groups:

yo18.29 <- CovidB[CovidB[, 4] == "18-29 years", ]
yo30.39 <- CovidB[CovidB[, 4] == "30-39 years", ]
yo40.49 <- CovidB[CovidB[, 4] == "40-49 years", ]
yo50.64 <- CovidB[CovidB[, 4] == "50-64 years", ]
yo65.74 <- CovidB[CovidB[, 4] == "65-74 years", ]
yo75.84 <- CovidB[CovidB[, 4] == "75-84 years", ]
yo85. <- CovidB[CovidB[, 4] == "85 years and over", ]
yo.18.85 <- rbind.data.frame(yo18.29, yo30.39, yo40.49, yo50.64,
    yo65.74, yo75.84, yo85.)

zee.table = xtabs(Death ~ Age_Group + Sex, data = yo.18.85, drop.unused.levels = TRUE)
# zee.table
the.test = chisq.test(zee.table, correct = FALSE)
eij = the.test$expected
# eij
chi.sq.obs = as.numeric(the.test$statistic)
# chi.sq.obs


# subsetting
# for
# categorical
# variables
# data
# frame

Age_Group = rep(c("18-29 years", "30-39 years", "40-49 years",
    "50-64 years", "65-74 years", "75-84 years", "85 years and older"),
    times = c(1684, 5028, 13482, 70152, 103439, 133552, 151341))
Sex = rep(c("Female", "Male", "Female", "Male", "Female", "Male",
    "Female", "Male", "Female", "Male", "Female", "Male", "Female",
    "Male"), times = c(640, 1044, 1736, 3292, 4582, 8900, 24893,
    45259, 40030, 63409, 58426, 75126, 87402, 63939))
data18.85 = data.frame(Age_Group, Sex)
# View(data18.85)
```

```r
# Plots:

# Interactions
# Plot

library(ggplot2)
library("plyr")
DeathSex <- ddply(yo.18.85, .(Age_Group, Sex), summarise, val = mean(Death))

ggplot(yo.18.85, aes(x = factor(Age_Group), y = Death, colour = Sex)) +
    geom_boxplot(aes(fill = Age_Group)) + geom_point(data = DeathSex,
    aes(y = val)) + geom_line(data = DeathSex, aes(y = val, group = Sex)) +
    theme_bw() + ggtitle("Interaction Plot of COVID-19 Death Counts Between Age and Sex",
    " n=478678") + theme(plot.title = element_text(hjust = 0.5)) +
    xlab("Age Groups") + ylab("Death Count") + labs(caption = "Figure 1: Variability of Death counts ap
        \t\t In any given Age-group other than the age groups of 85 and older,
        \t\t Males have a higher death count than Females.
         \t\tNumber of deaths per age group do not appear to be the same regardless of sex, and number
           Death counts are skewed right for Age-groups over the age of 65.") +
    theme(plot.caption = element_text(hjust = 0), axis.text.x = element_text(angle = 45,
        hjust = 1), legend.key.size = unit(0.35, "cm"), legend.key.width = unit(0.5,
        "cm"))  # Change legend key size and key width


# Grouped
# Bar
# Plot

library(ggplot2)
ggplot(yo.18.85, aes(fill = Age_Group, y = Death, x = Sex), color = Sex) +
    geom_bar(position = "dodge", stat = "identity", aes(fill = Age_Group)) +
    theme_bw() + ggtitle("Death Counts from COVID-19 Between Age and Sex",
    "n = 478678") + theme(plot.title = element_text(hjust = 0.5)) +
    xlab("Sex") + ylab("Death Count") + labs(caption = "Figure 2: Number of deaths per age group do not
    theme(plot.caption = element_text(hjust = 0))



# Summary
# Statistics:

# Expected
# count
# on
# average

eij

# Contingency
# table

zee.table
```

```r
ni. = rowSums(zee.table)
n.j = colSums(zee.table)
n = sum(zee.table)


CovidB <- read.csv("C:/Users/kulra/Desktop/CovidB.csv")
# Nonparametric
# Chi-squared
# Permutation
# Test
# For
# Independence
zee.table = table(data18.85)
the.test = chisq.test(zee.table, correct = FALSE)
chi.sq.obs = as.numeric(the.test$statistic)
# Fixed
# Row
# Totals
# And
# Randomly
# Assign
# Subjects
# In
# Row
# Into
# A
# Column

set.seed(104)
R = 3000
r.perms = sapply(1:R, function(i) {
    perm.data = data18.85
    perm.data$Age_Group = sample(perm.data$Age_Group, 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


# Fixed
# Column
# Totals
# And
# Randomly
# Assign
# Subjects
# In
# Column
# Into
# A
```

```r
# Row

set.seed(104)
R = 3000
r.perms = sapply(1:R, function(i) {
    perm.data = data18.85
    perm.data$Age_Group = sample(perm.data$Age_Group, 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

# Multiple
# Comparisons
# In
# Contingency
# Tables:
# A
# Test
# For
# Level
# Of
# Dependency

# Zij
# Of
# Pairwise
# Differences
# For
# P(Age
# |
# Sex)
# -P(Age'
# |
# Sex)

all.pjG1 = zee.table[1, ]/ni.[1]  #18-29
all.pjG1
all.pjG2 = zee.table[2, ]/ni.[2]  #30-39
all.pjG2
all.pjG3 = zee.table[3, ]/ni.[3]  #40-49
all.pjG3
all.pjG4 = zee.table[4, ]/ni.[4]  #50-64
all.pjG4
all.pjG5 = zee.table[5, ]/ni.[5]  #65-74
all.pjG5
all.pjG6 = zee.table[6, ]/ni.[6]  #75-84
all.pjG6
all.pjG7 = zee.table[7, ]/ni.[7]  #85+
all.pjG7
```

```
all.pbar = n.j/n  #all probabilities regardless of group
all.pbar
all.Zij1 = c(all.pjG1 - all.pjG2)/sqrt(all.pbar * (1 - all.pbar) *
    (1/ni.[1] + 1/ni.[2]))  #(18-29) - (30-39)
all.Zij1
all.Zij2 = c(all.pjG1 - all.pjG3)/sqrt(all.pbar * (1 - all.pbar) *
    (1/ni.[1] + 1/ni.[3]))  #(18-29) - (40-49)
all.Zij2
all.Zij3 = c(all.pjG1 - all.pjG4)/sqrt(all.pbar * (1 - all.pbar) *
    (1/ni.[1] + 1/ni.[4]))  #(18-29) - (50-64)
all.Zij3
all.Zij4 = c(all.pjG1 - all.pjG5)/sqrt(all.pbar * (1 - all.pbar) *
    (1/ni.[1] + 1/ni.[5]))  #(18-29) - (65-74)
all.Zij4
all.Zij5 = c(all.pjG1 - all.pjG6)/sqrt(all.pbar * (1 - all.pbar) *
    (1/ni.[1] + 1/ni.[6]))  #(18-29) - (75-84)
all.Zij5
all.Zij6 = c(all.pjG1 - all.pjG7)/sqrt(all.pbar * (1 - all.pbar) *
    (1/ni.[1] + 1/ni.[7]))  #(18-29) - (85+)
all.Zij6
all.Zij7 = c(all.pjG2 - all.pjG3)/sqrt(all.pbar * (1 - all.pbar) *
    (1/ni.[2] + 1/ni.[3]))  #(30-39) - (40-49)
all.Zij7
all.Zij8 = c(all.pjG2 - all.pjG4)/sqrt(all.pbar * (1 - all.pbar) *
    (1/ni.[2] + 1/ni.[4]))  #(30-39) - (50-64)
all.Zij8
all.Zij9 = c(all.pjG2 - all.pjG5)/sqrt(all.pbar * (1 - all.pbar) *
    (1/ni.[2] + 1/ni.[5]))  #(30-39) - (65-74)
all.Zij9
all.Zij10 = c(all.pjG2 - all.pjG6)/sqrt(all.pbar * (1 - all.pbar) *
    (1/ni.[2] + 1/ni.[6]))  #(30-39) - (75-84)
all.Zij10
all.Zij11 = c(all.pjG2 - all.pjG7)/sqrt(all.pbar * (1 - all.pbar) *
    (1/ni.[2] + 1/ni.[7]))  #(30-39) - (85+)
all.Zij11
all.Zij12 = c(all.pjG3 - all.pjG4)/sqrt(all.pbar * (1 - all.pbar) *
    (1/ni.[3] + 1/ni.[4]))  #(40-49) - (50-64)
all.Zij12
all.Zij13 = c(all.pjG3 - all.pjG5)/sqrt(all.pbar * (1 - all.pbar) *
    (1/ni.[3] + 1/ni.[5]))  #(40-49) - (65-74)
all.Zij13
all.Zij14 = c(all.pjG3 - all.pjG6)/sqrt(all.pbar * (1 - all.pbar) *
    (1/ni.[3] + 1/ni.[6]))  #(40-49) - (75-84)
all.Zij14
all.Zij15 = c(all.pjG3 - all.pjG7)/sqrt(all.pbar * (1 - all.pbar) *
    (1/ni.[3] + 1/ni.[7]))  #(40-49) - (85+)
all.Zij15
all.Zij16 = c(all.pjG4 - all.pjG5)/sqrt(all.pbar * (1 - all.pbar) *
    (1/ni.[4] + 1/ni.[5]))  #(50-64) - (65-74)
all.Zij16
all.Zij17 = c(all.pjG4 - all.pjG6)/sqrt(all.pbar * (1 - all.pbar) *
    (1/ni.[4] + 1/ni.[6]))  #(50-64) - (75-84)
all.Zij17
```

```
all.Zij18 = c(all.pjG4 - all.pjG7)/sqrt(all.pbar * (1 - all.pbar) *
    (1/ni.[4] + 1/ni.[7]))  #(50-64) - (85+)
all.Zij18
all.Zij19 = c(all.pjG5 - all.pjG6)/sqrt(all.pbar * (1 - all.pbar) *
    (1/ni.[5] + 1/ni.[6]))  #(65-74) - (75-84)
all.Zij19
all.Zij20 = c(all.pjG5 - all.pjG7)/sqrt(all.pbar * (1 - all.pbar) *
    (1/ni.[5] + 1/ni.[7]))  #(65-74) - (85+)
all.Zij20
all.Zij21 = c(all.pjG6 - all.pjG7)/sqrt(all.pbar * (1 - all.pbar) *
    (1/ni.[6] + 1/ni.[7]))  #(75-84) - (85+)
all.Zij21

# Finding
# Tukey's
# HSD
# Cutoff
# Values
# For
# P(Age
# |
# Sex)
# -P(Age'
# |
# Sex)

set.seed(104)
R = 2000
r.perms.cutoff = sapply(1:R, function(i) {
    perm.data = data18.85
    perm.data$Age_Group = sample(perm.data$Age_Group, nrow(perm.data),
        replace = FALSE)
    row.sum = rowSums(table(perm.data))
    col.sum = colSums(table(perm.data))
    all.pji1 = table(perm.data)[1, ]/row.sum[1]
    all.pji2 = table(perm.data)[2, ]/row.sum[2]
    all.pji3 = table(perm.data)[3, ]/row.sum[3]
    all.pji4 = table(perm.data)[4, ]/row.sum[4]
    all.pji5 = table(perm.data)[5, ]/row.sum[5]
    all.pji6 = table(perm.data)[6, ]/row.sum[6]
    all.pji7 = table(perm.data)[7, ]/row.sum[7]
    all.pbar = col.sum/sum(row.sum)
    all.Zij = c((all.pji1 - all.pji2)/sqrt(all.pbar * (1 - all.pbar) *
        (1/row.sum[1] + 1/row.sum[2])), (all.pji1 - all.pji3)/sqrt(all.pbar *
        (1 - all.pbar) * (1/row.sum[1] + 1/row.sum[3])), (all.pji1 -
        all.pji4)/sqrt(all.pbar * (1 - all.pbar) * (1/row.sum[1] +
        1/row.sum[4])), (all.pji1 - all.pji5)/sqrt(all.pbar *
        (1 - all.pbar) * (1/row.sum[1] + 1/row.sum[5])), (all.pji1 -
        all.pji6)/sqrt(all.pbar * (1 - all.pbar) * (1/row.sum[1] +
        1/row.sum[6])), (all.pji1 - all.pji7)/sqrt(all.pbar *
        (1 - all.pbar) * (1/row.sum[1] + 1/row.sum[7])), (all.pji2 -
        all.pji3)/sqrt(all.pbar * (1 - all.pbar) * (1/row.sum[2] +
        1/row.sum[3])), (all.pji2 - all.pji4)/sqrt(all.pbar *
```

```
        (1 - all.pbar) * (1/row.sum[2] + 1/row.sum[4])), (all.pji2 -
        all.pji5)/sqrt(all.pbar * (1 - all.pbar) * (1/row.sum[2] +
        1/row.sum[5])), (all.pji2 - all.pji6)/sqrt(all.pbar *
        (1 - all.pbar) * (1/row.sum[2] + 1/row.sum[6])), (all.pji2 -
        all.pji7)/sqrt(all.pbar * (1 - all.pbar) * (1/row.sum[2] +
        1/row.sum[7])), (all.pji3 - all.pji4)/sqrt(all.pbar *
        (1 - all.pbar) * (1/row.sum[3] + 1/row.sum[4])), (all.pji3 -
        all.pji5)/sqrt(all.pbar * (1 - all.pbar) * (1/row.sum[3] +
        1/row.sum[5])), (all.pji3 - all.pji6)/sqrt(all.pbar *
        (1 - all.pbar) * (1/row.sum[3] + 1/row.sum[6])), (all.pji3 -
        all.pji7)/sqrt(all.pbar * (1 - all.pbar) * (1/row.sum[3] +
        1/row.sum[7])), (all.pji4 - all.pji5)/sqrt(all.pbar *
        (1 - all.pbar) * (1/row.sum[4] + 1/row.sum[5])), (all.pji4 -
        all.pji6)/sqrt(all.pbar * (1 - all.pbar) * (1/row.sum[4] +
        1/row.sum[6])), (all.pji4 - all.pji7)/sqrt(all.pbar *
        (1 - all.pbar) * (1/row.sum[4] + 1/row.sum[7])), (all.pji5 -
        all.pji6)/sqrt(all.pbar * (1 - all.pbar) * (1/row.sum[5] +
        1/row.sum[6])), (all.pji5 - all.pji7)/sqrt(all.pbar *
        (1 - all.pbar) * (1/row.sum[5] + 1/row.sum[7])), (all.pji6 -
        all.pji7)/sqrt(all.pbar * (1 - all.pbar) * (1/row.sum[6] +
        1/row.sum[7])))
    Q.r = max(abs(all.Zij))
    return(Q.r)
})
hist(r.perms.cutoff)
alpha = 0.05
cutoff.q = as.numeric(quantile(r.perms.cutoff, (1 - alpha)))


# Zij
# Of
# Pairwise
# Differences
# For
# P(Sex
# |
# Age)
# -P(Sex'
# |
# Age)

zee.table = table(data18.85)
n = sum(zee.table)
ni. = rowSums(zee.table)
n.j = colSums(zee.table)
all.pjG3 = zee.table[, 1]/n.j[1]  #all conditional probabilites for row 1
all.pjG4 = zee.table[, 2]/n.j[2]  #all conditional probabilites for row 2
all.pbar2 = ni./n  #all probabilities regardless of group
all.Zij22 = c((all.pjG3 - all.pjG4)/sqrt(all.pbar2 * (1 - all.pbar2) *
    (1/(n.j[1]) + 1/(n.j[2]))))
all.Zij22
all.Zij22 = matrix(all.Zij22, nrow = 1)
colnames(all.Zij22) = c("18-29 years", "30-39 years", "40-49 years",
```

```
    "50-64 years", "65-74 years", "75-84 years", "85 years and older")
rownames(all.Zij22) = c("Female vs. Male")
all.Zij22

# Finding
# Tukey's
# HSD
# Cutoff
# Values
# For
# P(Female
# |
# Age)
# -
# P(Male
# |
# Age)

set.seed(104)
R = 2000
r.perms.cutoff1 = sapply(1:R, function(i) {
    perm.data = data18.85
    perm.data$Sex = sample(perm.data$Sex, nrow(perm.data), replace = FALSE)
    row.sum = rowSums(table(perm.data))
    col.sum = colSums(table(perm.data))
    all.pji8 = table(perm.data)[, 1]/col.sum[1]
    all.pji9 = table(perm.data)[, 2]/col.sum[2]
    all.pbar1 = row.sum/sum(col.sum)
    all.Zij3 = c((all.pji8 - all.pji9)/sqrt(all.pbar1 * (1 -
        all.pbar1) * (1/col.sum[1] + 1/col.sum[2])))
    Q.r = max(abs(all.Zij3))
    return(Q.r)
})
hist(r.perms.cutoff1)
alpha = 0.05
cutoff.q1 = as.numeric(quantile(r.perms.cutoff1, (1 - alpha)))
cutoff.q1

# All
# Pairwise
# Zij
# for
# Comparison
# of
# Ages
# Given
# Sex
# Tukey's
# HSD
# Cutoff
# for
# Comparison
# of
```

```
# Ages
# Given
# Sex
# Histogram
# of
# Tukey's
# HSD
# cutoff
# permutation
# distribution
# for
# Comparison
# of
# Ages
# Given
# Sex
# All
# Pairwise
# Zij
# for
# Comparison
# of
# Sexes
# Given
# Age
# Tukey's
# HSD
# Cutoff
# for
# Comparison
# of
# Sexes
# Given
# Age
# Histogram
# of
# Tukey's
# HSD
# cutoff
# permutation
# distribution
# for
# Comparison
# of
# Sexes
# Given
# Age

all.Zij
cutoff.q
hist(r.perms.cutoff)
all.Zij22
cutoff.q1
```

```r
hist(r.perms.cutoff1)
```