Statistical inference with the General Social Survey dataset

1. The context

The data is obtained from the General Social Survey (GSS) dataset. The GSS gathers data regarding demographic, behavioral, attitudinal aspects, and other topics of special interest in American society in order to monitor and explain trends and constants in behaviors, attitudes, and attributes.

This project has 4 objectives:

- **Data**: How are the observations in this GSS sample collected? How does this data collection method imply on the scope of inference in terms of generalizability & causality?
- **Research question:** Come up with a research question based on this dataset.
- **EDA**: Perform exploratory data analysis addressing the research question outlined.
- **Inference**: Perform statistical inference (hypothesis testing) addressing the research question stated.

SET UP IN R
Load packages

```
library(ggplot2)
library(dplyr)
library(statsr)
```

Load data

```
load("gss.Rdata")
```

2. Data

Observations in the sample are collected based on simple random sampling method. This is because the GSS sample respondents are randomly selected in households across the United States. Respondents are mixed between various geographical areas from rural, suburban, and urban areas.

Data from this sample are 57,061 randomly people with 114 variables. The sample size is large enough to avoid annecdotal evidence issue. As discussed above, sample data are collected in a non-biasd way. Therefore, we can conclude that random sampling was used and results from this sample are acceptable to be generalized to a larger population.

As survey randomly collected data from the US demographics, there is no sign of random assignment being used. However, even when the researchers implement random assignment, it

could be quite challenging to implement as there are potential unexpected variables which locate outside the laboratory conditions.

In this case, we can use inductive reasoning to infer causality. There are potential variables can be used to infer causality (confinan: confidence in banks and financial institutions, confed: confidence in executive branch of federal government, etc.). To infer causality, these data must be exogenous. However, these ratings are most likley endogenous and caused by, say, political party affiliation or think of self as liberal or conservative. There will be a lot of possible cofounding factors. Therefore, a causality study would be hard.

3. Research question

Controlling for observable characteristics, does the average family income (in constant dollars) is the same across all races (White, Black, Other)?

The reason why this question is interesting is that this research surveys across a broad range of US demographics, including all races. Studying the differences between average family income across races may reveal helpful insights about the income gap in the US.

4. Exploratory data analysis

First, we will start with exploring different facets of our family income data. We start with the measure of center, specifically the mean. We want to find out what is the average family income of people conducting this survey. We exclude NA answers in our computation.

```
mean(gss$coninc, trim = 0, na.rm = TRUE)
```

```
## [1] 44503.04
```

The result shows that the average family income is 44,503.04 (rounded to 44,504) US dollars.

Next, we will examine the median value. Again, we exclude NA answers in our computation.

```
median(gss$coninc, trim = 0, na.rm = TRUE)
```

```
## [1] 35602
```

The median value is 35,602 US dollars. There is quite a big difference between the mean and the median values.

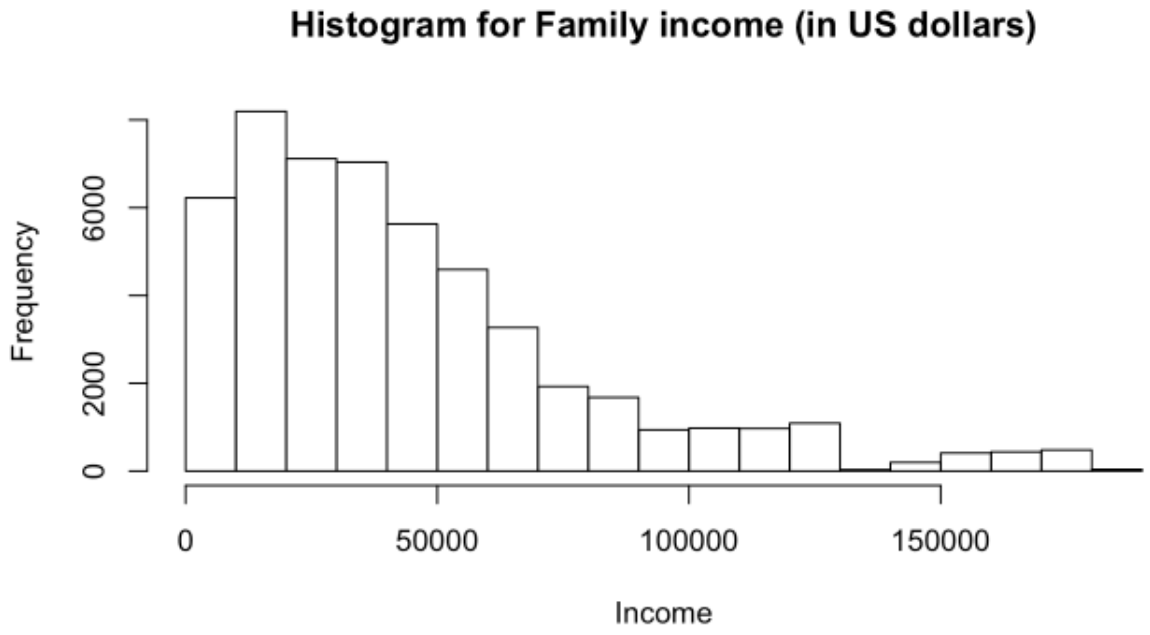
Coming to the measure of spread, we will calculate the standard deviation.

```
sd(gss$coninc, na.rm = TRUE)
## [1] 35936.01
```

The standard deviation is 35,936.01. The data is quite spread around the mean.

Based on above analysis, we can have a more throughout view based on a histogram.

```
hist(gss$coninc, main= "Histogram for Family income (in US dollars)",
 xlab = "Income")
```



The histogram shows that data are unimodal. The data distribution is heavily right-skewed.

Another way to measure spread is interquartile range.

```
quantile(gss$coninc, na.rm = TRUE)
##     0%    25%    50%    75%   100%
##    383  18445  35602  59542 180386
```

The first, second, and third quartiles of coninc is 18,445 – 35,602 – 44,503 respectively.

We can visualize in a box plot:

```
boxplot(gss$coninc, data=gss, notch=FALSE, col=(c("gold","darkgreen"))
```

```
,main="Box plot for Family income", xlab="Income")
```

The interquartile range:

IQR = Q3 – Q1 = 44,503 – 18,445 = 26,058.

To sum up our exploratory data analysis, we can do summary statistics:

```
summary(gss$coninc)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##     383   18445   35602   44503   59542  180386    5829
```

That concludes all of our above analysis regarding exploratory data analysis.

5. Inference

As we are comparing the differences in mean (more than 2 means), the method we use is ANOVA.

Set the hypotheses

μ = average family income (in US dollars)

H0: The average family income is the same across all races.

μ1 = μ2 = μ3

HA: The average family income differs between at least one pair of races.

Check conditions

Independence

Within groups
As mentioned above, samples are collected based on random sampling method. The sample size (57,061) is less than 10% of the population (population of the US). Therefore, the condition of independence between groups is guaranteed.

Between groups
As this is individual survey, dependence between groups is not likely to happen. For instance, people with different races or education levels do not have any sign to be dependent to each other.

We can conclude that the independence conditions (both within and between groups) are ensured.

Approximate normality

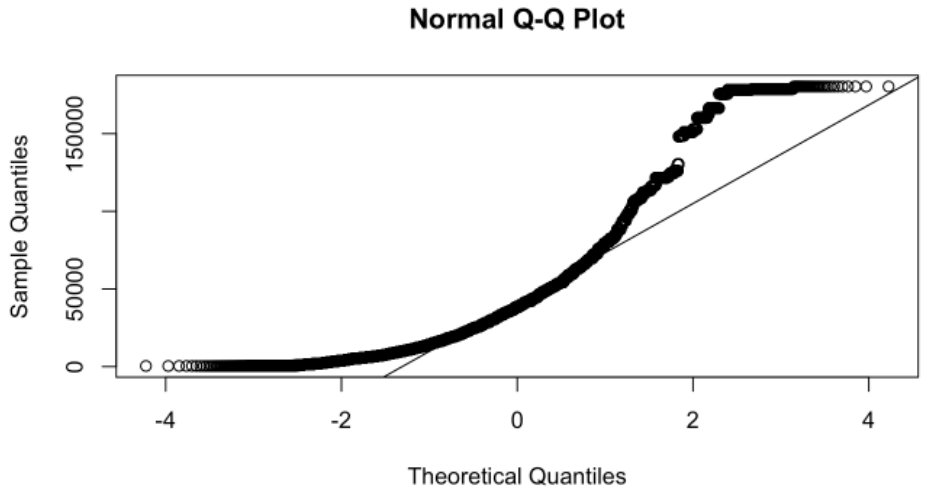
To verify this condition, we draw normal Q-Q plots.

First, we start with the White race by creating a subset:

```
white <- subset(gss, race == "White")
```

Let's see a Q-Q plot:

```
qqnorm(white$coninc)
qqline(white$coninc)
```



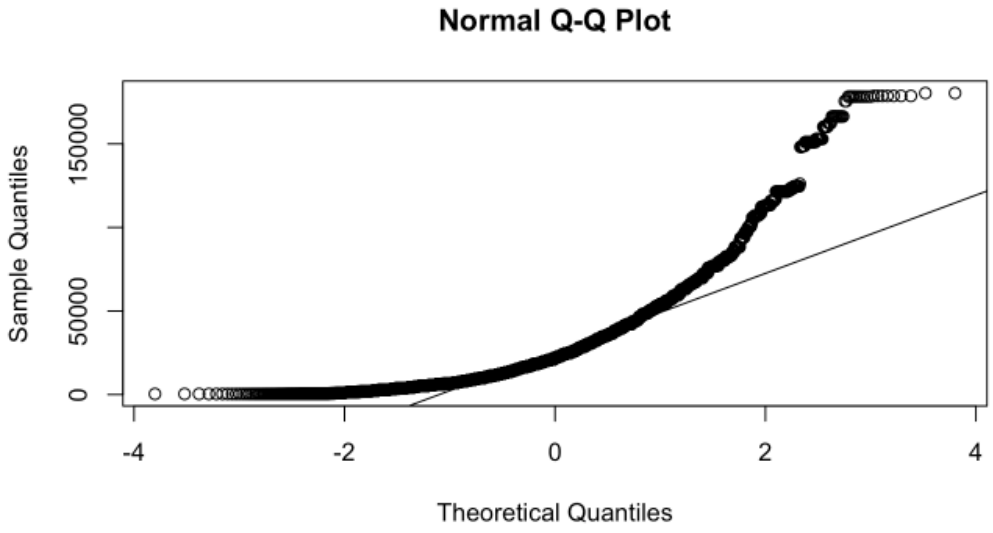
We can see that distribution of response variable (White group) looks approximately normal.

Second, we will create a subset with the Black group:

```
black <- subset (gss, race == "Black")
```

Let's see a Q-Q plot:

```
qqnorm(black$coninc)
qqline(black$coninc)
```



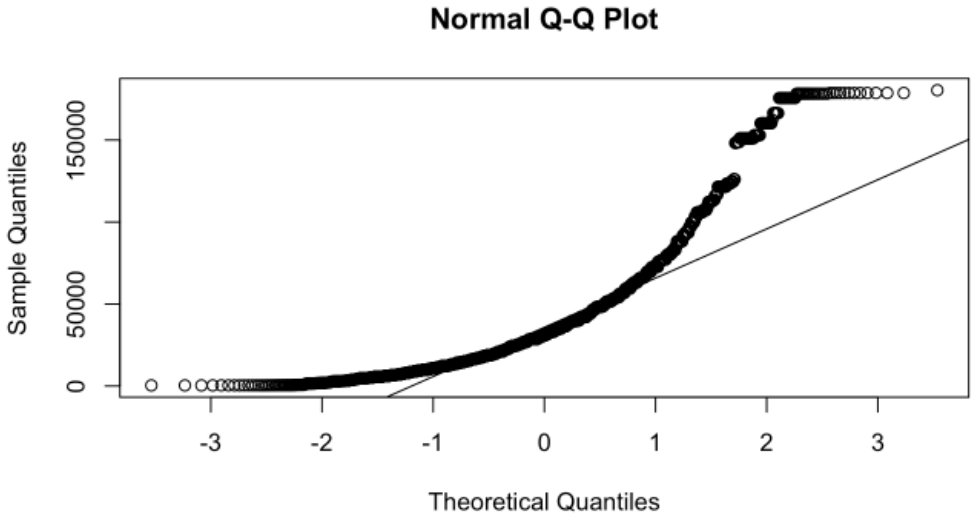
The distribution looks not really normal, but it is acceptable.

Finally, we create a subset with the Other group:

```
other <- subset(gss, race == "Other")
```

Let's see a Q-Q plot:

```
qqnorm(other$coninc)
qqline(other$coninc)
```



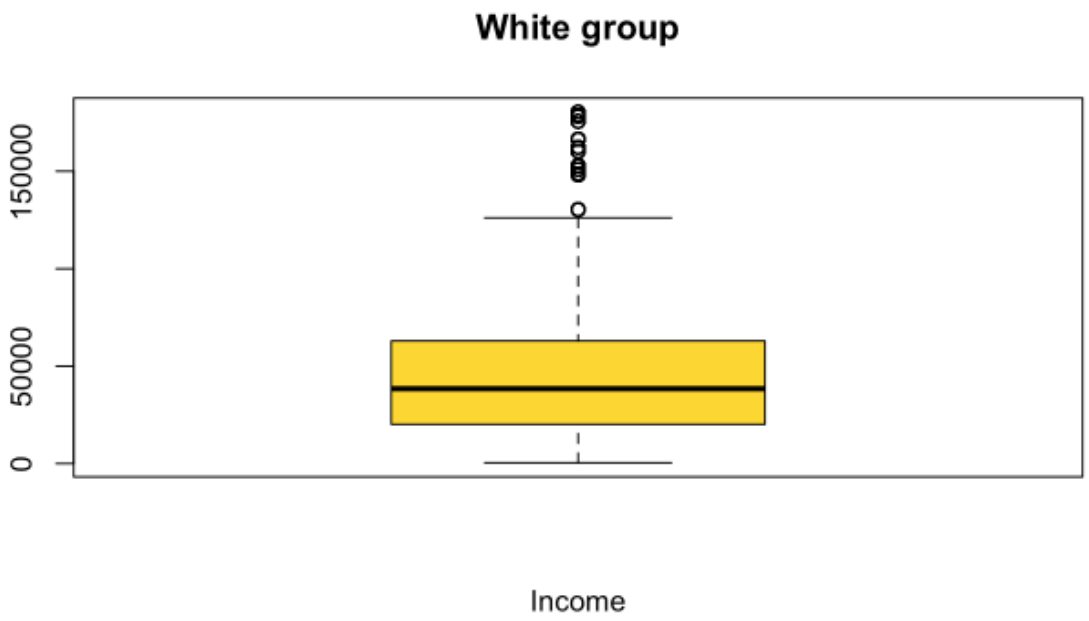
The distribution of the Other group looks quite the same with the distribution of the Black group.

Based on the above Q-Q plots, we can conclude that the second condition is quite met.

Equal variance

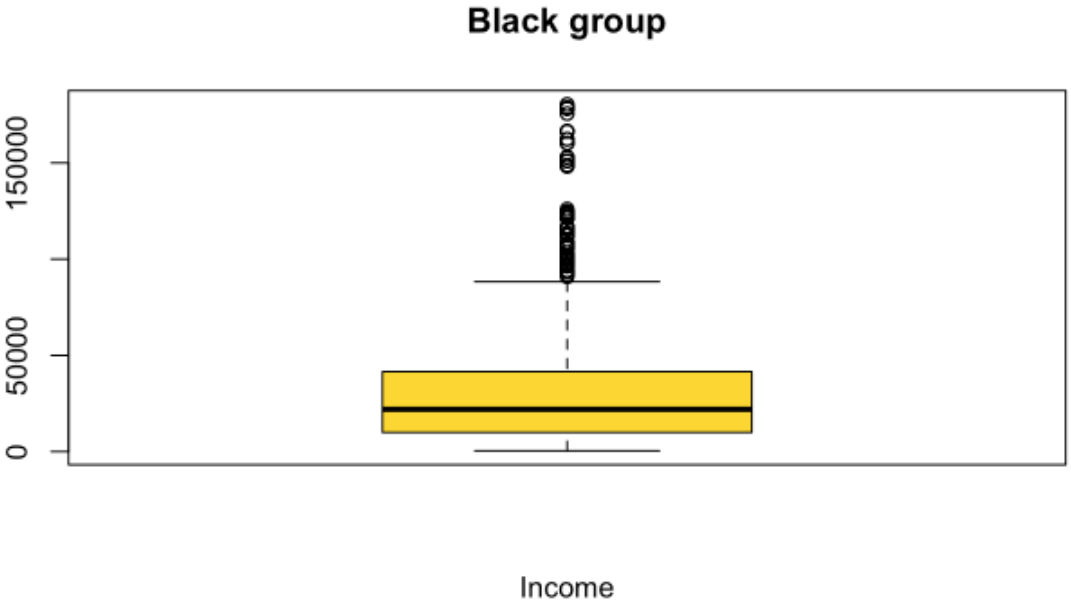
We will draw boxplots for different groups. First, we start with the White group.

```
boxplot(white$coninc, data=white, notch=FALSE, col=(c("gold","darkgreen
", main="White group", xlab="Income")
```



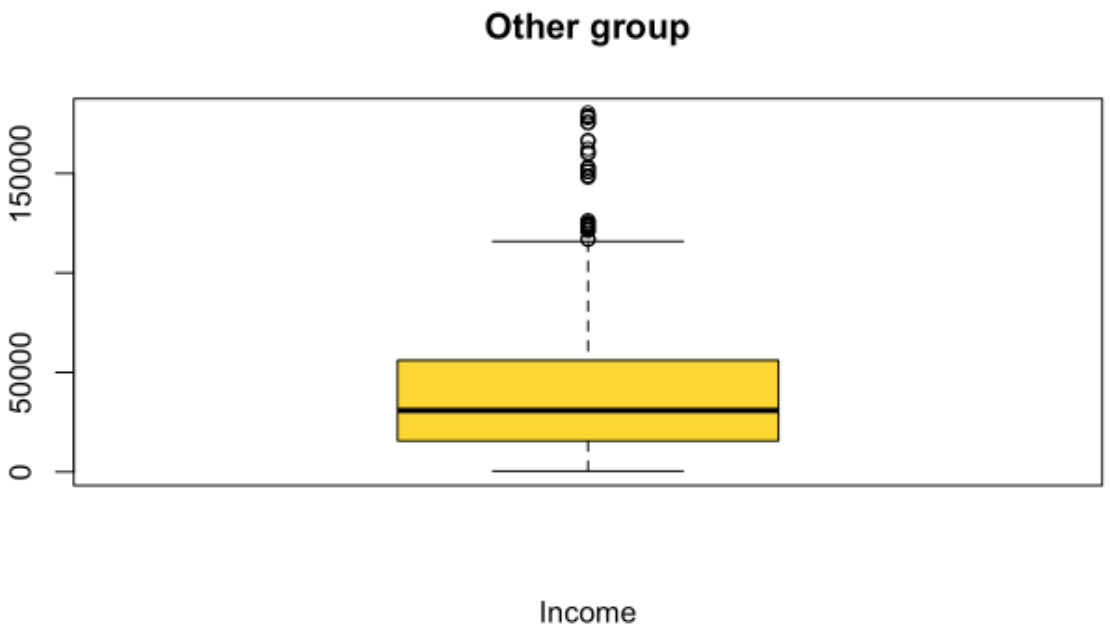
Second, we draw a boxplot for the Black group.

```
boxplot(black$coninc, data=black, notch=FALSE
, col=(c("gold","darkgreen"))
, main="Black group", xlab="Income")
```



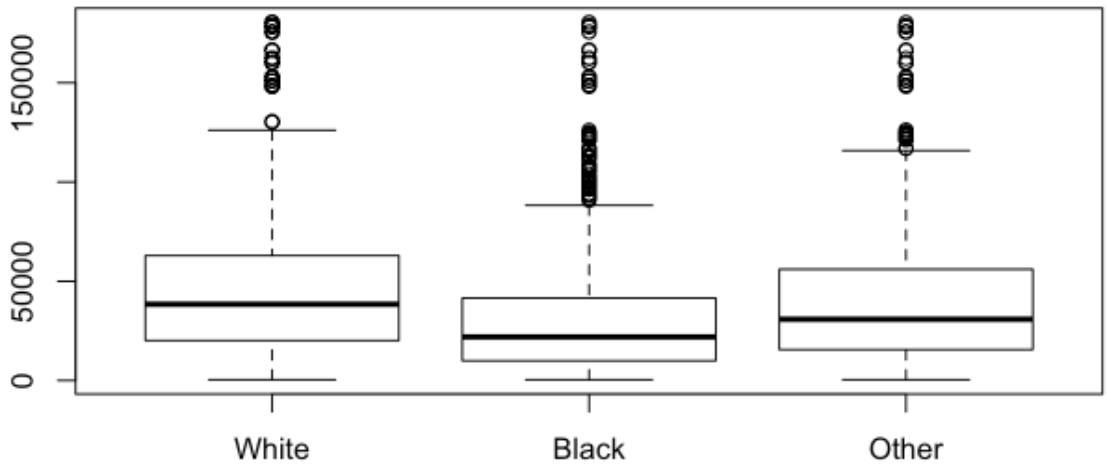
Finally, we draw a box plot for the Other group.

```
boxplot(other$coninc, data=other, notch=FALSE
, col=(c("gold","darkgreen"))
, main="Other group", xlab="Income")
```



In comparison between 3 groups:

```
boxplot(gss$coninc ~ gss$race)
```



From the 3 box plots, we see that variability is constant across 3 groups. Therefore, the constant variance condition is met.

Methods stated and described

We will use a one-way ANOVA test in this case. This is because we have 1 independent variable (family income in US dollars). We want to compare the differences in the average family income between 3 races (White, Black, Other). The number of observations is not the same between those 3 groups.

How to perform the test? Using the aov function, we can observe the sum of squares total (SST), sum of squares group (SSG), sum of squares error (SSE), degrees of freedom, mean squares, F statistic, and p-value.

Perform inference

Let's fit the model by the ANOVA formula:

```
1   aov.out = aov(coninc ~ race, data=gss)
2   summary(aov.out)
```

```
##                Df    Sum Sq   Mean Sq F value Pr(>F)
## race            2 1.699e+12 8.494e+11   675.1 <2e-16 ***
## Residuals   51229 6.446e+13 1.258e+09
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 5829 observations deleted due to missingness
```

We can have a more throughout analysis based on the lm function:

```
income_race = lm(gss$coninc ~ gss$race -1, data = gss)
summary(income_race)
```

```
##
## Call:
## lm(formula = gss$coninc ~ gss$race - 1, data = gss)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -46624 -25028  -8593  14714 150201
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## gss$raceWhite  47006.7      173.5  271.01   <2e-16 ***
## gss$raceBlack  30185.0      425.3   70.97   <2e-16 ***
## gss$raceOther  42415.4      716.4   59.21   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 35470 on 51229 degrees of freedom
##   (5829 observations deleted due to missingness)
## Multiple R-squared:  0.6154, Adjusted R-squared:  0.6154
## F-statistic: 2.733e+04 on 3 and 51229 DF,  p-value: < 2.2e-16
```

We can also calculate the 95% confidence interval:

```
confint(income_race,level=.95)
```

```
##                      2.5 %   97.5 %
## gss$raceWhite 46666.78 47346.71
## gss$raceBlack 29351.40 31018.64
## gss$raceOther 41011.36 43819.49
```

Interpretations and conclusions

From the ANOVA and lm results, we observe that it is highly significant as the corresponding p-value is really small. The p-value is < 0.00000000000000022, which is much smaller than the conventional value 0.05.

In conclusion, because the p-value is small, we reject the null hypothesis. The data provide convincing evidence that at least one pair of population means are different from each other. There is a difference between the average family income (is US dollars) between at least one pair of races.

Reasoning for why CI is/is not also included

The 95% confidence interval indicates that you can be 95% confident that all the confidence intervals contain the true differences.

Based on the confidence interval results, we are 95% confident that White people's family income on average range from 46666.78 to 47346.71 US dollars. We are 95% confident that Black people's family income on average range from 29351.40 to 31018.64 US dollars. We are 95% confident that Other-race people's family income on average range from 41011.36 to 43819.49 US dollars.

We can observe that the CI range does not include zero, indicating that the difference is statistically significant. If the CI range of any pair includes zero, the differences are not statistically significant.

***