Statistical Method Illustration - Testing

Daniel Shang

```r
# Load the necessary packages
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(ggthemes)
```

```
## Warning: package 'ggthemes' was built under R version 4.0.3
```

```r
library(car)
```

```
## Warning: package 'car' was built under R version 4.0.3

## Loading required package: carData

##
## Attaching package: 'car'

## The following object is masked from 'package:dplyr':
##
##     recode
```

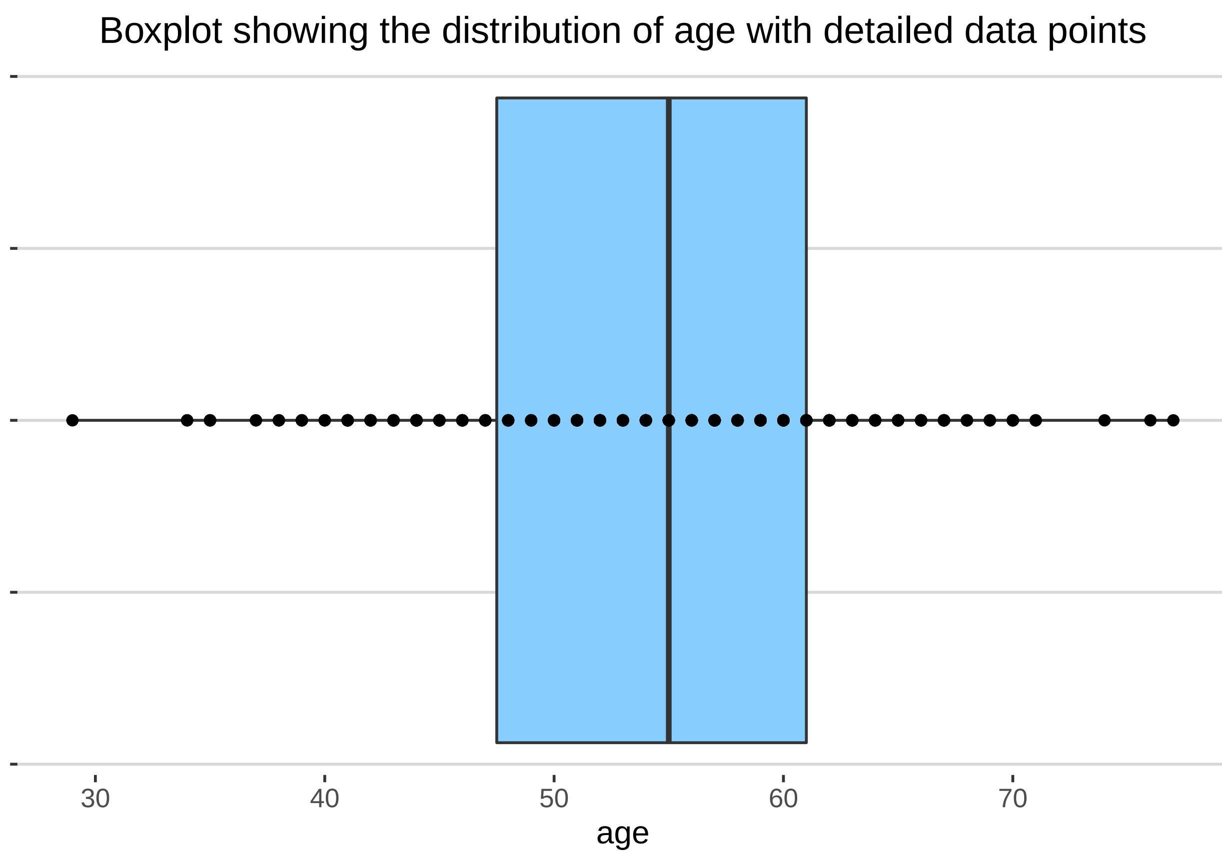
```r
library(ggplot2)
```

```r
# Load the data, remove missing values (if any), and convert columns to proper formats
## based on the documentation of the dataset
data = read.csv('heart.csv', )
data_clean = na.omit(mutate_all(data,
                      ~ifelse(. %in% c("N/A", "null", "", NULL),  NA, .)))
colnames(data_clean)[1] = 'age'
data_clean$sex = as.factor(data_clean$sex)
```

```
data_clean$cp = as.factor(data_clean$cp)
data_clean$fbs = as.factor(data_clean$fbs)
data_clean$restecg = as.factor(data_clean$restecg)
data_clean$exang = as.factor(data_clean$exang)
data_clean$slope = as.factor(data_clean$slope)
data_clean$ca = as.factor(data_clean$ca)
data_clean$thal = as.factor(data_clean$thal)
data_clean$target = as.factor(data_clean$target)
```

————————- Parametric Statistical Test (One-Samples T-Test) ————————-

```
# Make a boxplot to get a basic understanding of the distribution of the data.

## Please note that the example is not strict with the assumptions and conditions
## required by the test. The sole purpose of the illustration is to show how to conduct
## the test in terms of coding. Assume that they the data set(s) meet all the necessary
## assumptions and requirements of the test to be conducted.

ggplot(data = data_clean, aes(x = age)) +
  geom_boxplot(fill = 'skyblue1') +
  geom_point(aes(y = 0)) +
  theme_hc() +
  theme(axis.text.y = element_blank(), axis.title.y = element_blank(),
        legend.position = 'none', plot.title = element_text(hjust = 0.5)) +
  labs(title = 'Boxplot showing the distribution of age with detailed data points')
```



```r
# Calculate descriptive statistics
mean(data_clean$age)
```

```
## [1] 54.36634
```

```r
sd(data_clean$age)
```

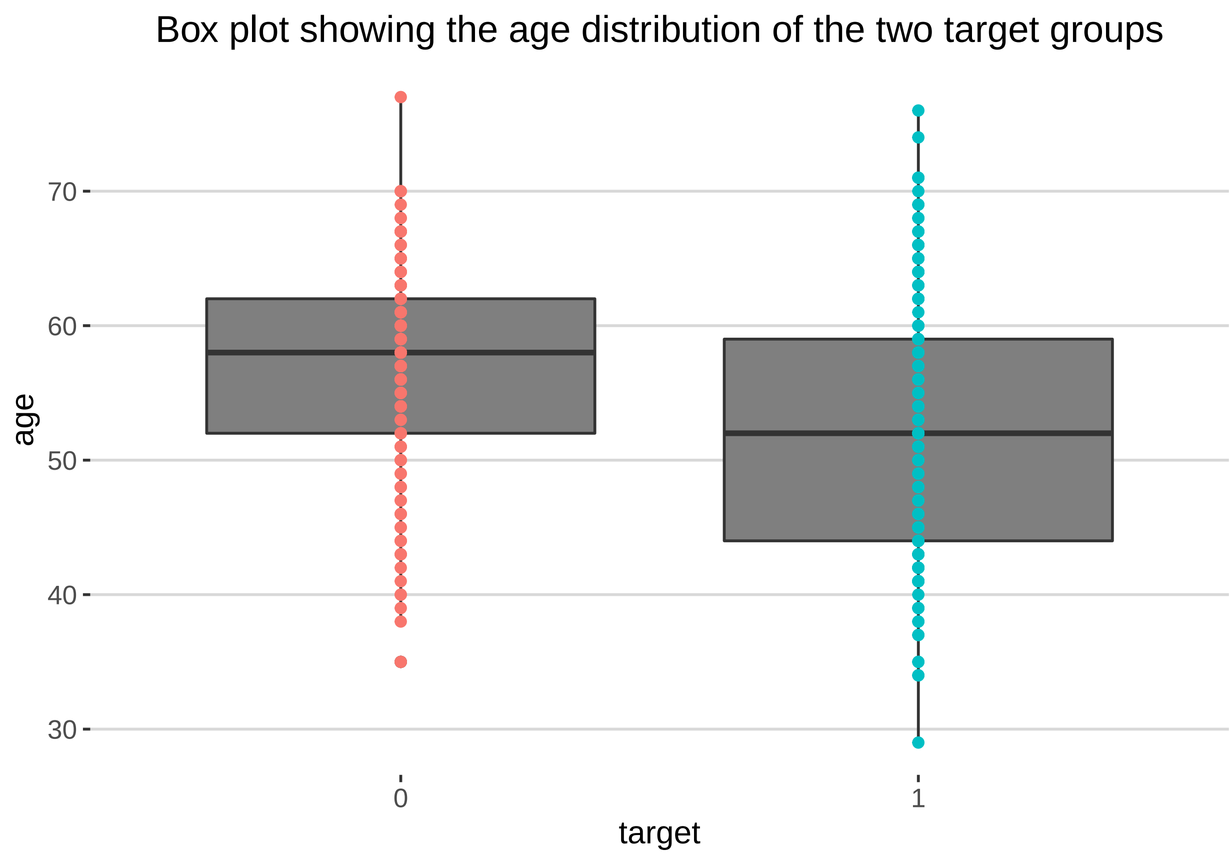
```
## [1] 9.082101
```

```r
# For this example, I will test if the average age of the population is less than
## 53 with 99% confidence level. With these, our null hypothesis is H0: mu = 53,
## and our alternative hypothesis is H1: mu != 53. Since the result shows a p-value
## of 0.009271, we can reject the null hypothesis with 99% confidence. Therefore,
## we are 99% confident that the average age of the population is not equal to 53.
t.test(data_clean$age, mu = 53, alternative = 'two.sided', conf.level = 0.99)
```

```
##
##  One Sample t-test
##
## data:  data_clean$age
## t = 2.6187, df = 302, p-value = 0.009271
## alternative hypothesis: true mean is not equal to 53
## 99 percent confidence interval:
##  53.01384 55.71883
## sample estimates:
## mean of x
##  54.36634
```

—————————- Parametric Statistical Test (Two-Sample T-Test) —————————-

```r
# In this example, I will conduct a independent 2-sample t-test. I will check whether
## the people in the two 'target' groups have the same age on average. With these,
## our null hypothesis is H0: mean age of target 0 group = mean age of target 1 group.
## Thus, the alternative hypothesis H1: the mean ages of the two groups differ.

## Please note that the example is not strict with the assumptions and conditions
## required by the test. The sole purpose of the illustration is to show how to conduct
## the test in terms of coding. Assume that they the data set(s) meet all the necessary
## assumptions and requirements of the test to be conducted.

ggplot(data = data_clean, aes(x = target, y = age)) +
  geom_boxplot(fill = 'gray50') +
  geom_point(aes(color = target)) +
  theme_hc() +
  labs(title = 'Box plot showing the age distribution of the two target groups') +
  theme(plot.title = element_text(hjust = 0.5), legend.position = 'none')
```




```r
# Since we can see from the box plot that target 0 group has a larger variation than
## the target 1 group, we will pass in argument that says the variances are not equal.
## I also passed in the argument saying that the two groups are independent from each
## other. The test result shows that the p-value is 5.781*10^-5. Therefore, we can
## conclude with 99% confidence that the average ages of the two target groups are
## different.
t.test(data_clean$age~data_clean$target, mu = 0, alternative = 'two.sided',
       conf.level = 0.99, var.eq = FALSE, paired = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  data_clean$age by data_clean$target
## t = 4.0797, df = 301, p-value = 5.781e-05
## alternative hypothesis: true difference in means is not equal to 0
## 99 percent confidence interval:
##  1.496453 6.712506
## sample estimates:
## mean in group 0 mean in group 1
##        56.60145        52.49697
```

```r
# To double check if the variances of the two target groups are indeed different, we
## can either calculate the variances of the two groups or conduct a Levene's test,
## which tests if the variances of datasets are equal. Since the variances are
## different and the test result shows a small p-value, we can reject the null
```

```r
## hypothesis that the two groups' variances are equal.
var(data_clean$age[data_clean$target == 0])
```

```
## [1] 63.39474
```

```r
var(data_clean$age[data_clean$target == 1])
```

```
## [1] 91.21493
```

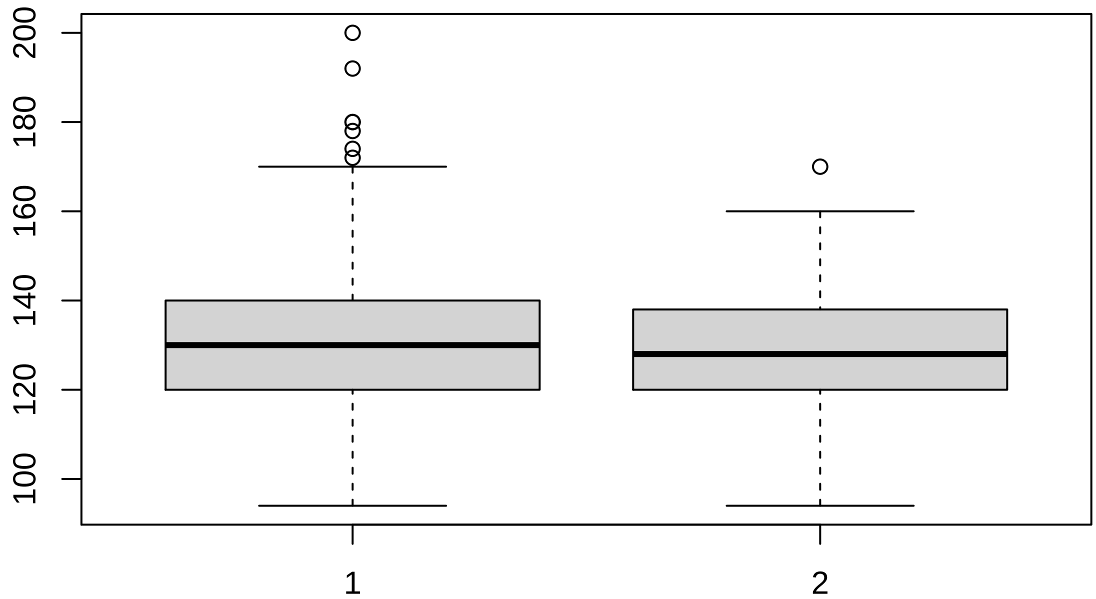
```r
leveneTest(data_clean$age~data_clean$target)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value   Pr(>F)
## group   1  7.9854 0.005031 **
##       301
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

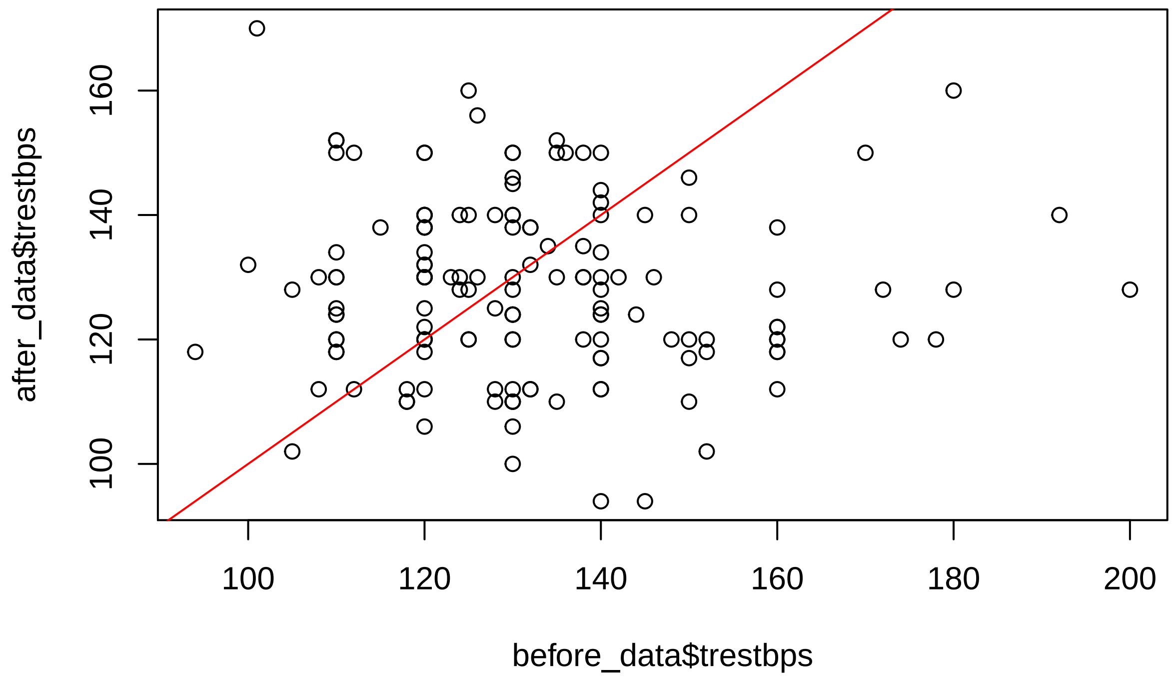
Parametric Statistical Test (Paired T-Test)

```r
# In this example, I will conduct a paired t-test to examine the difference in
## means of the two population data sets. Since this example is for illustration
## purpose only, I randomly split the data into two groups pair them together

## Please note that the example is not strict with the assumptions and conditions
## required by the test. The sole purpose of the illustration is to show how to conduct
## the test in terms of coding. Assume that they the data set(s) meet all the necessary
## assumptions and requirements of the test to be conducted.

set.seed(123)
index_1 = sample(2, nrow(data_clean), replace = TRUE, prob = c(0.5, 0.5))
before_data = data_clean[index_1 == 1, ]
after_data = data_clean[index_1 == 2, ]
index_2 = sample(147, nrow(after_data), replace = TRUE)
after_data = after_data[index_2[1:147], ]
```

```r
# Make a boxplot the to visualize the distributions of the two groups
boxplot(before_data$trestbps, after_data$trestbps)
```




```r
# Make a scatter plot to further visualize the relationship between the pair.
## If there is no significant difference between the means of the groups, the points
## should be evenly distributed on each side of the red line, which is a line with
## an intercept of 0 and slope of 1 for reference purpose. Here, we see that relatively
## more points are distributed below the line, so the means of the groups may not be
## equal.
plot(before_data$trestbps, after_data$trestbps, )
abline(a = 0, b = 1, col = 'red')
```



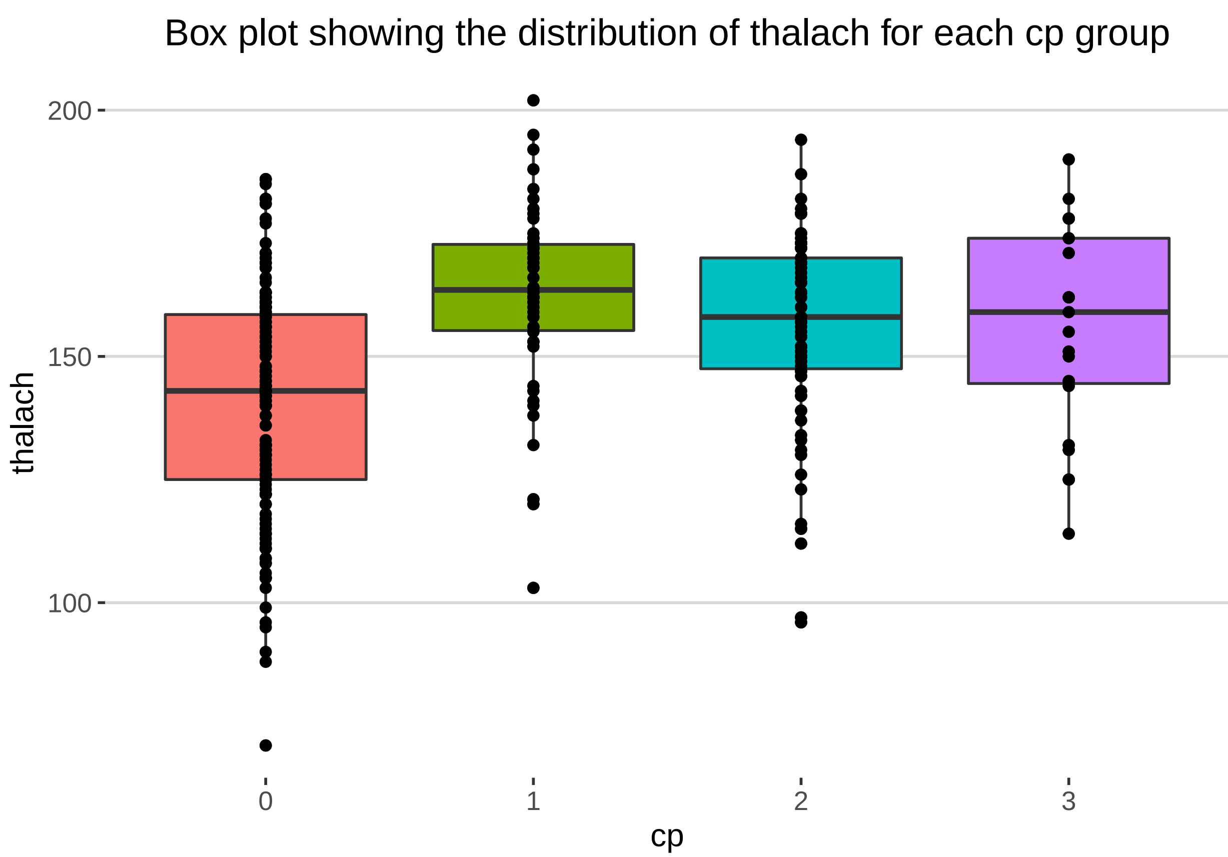

```r
# In this example, our null hypothesis is that the means for two groups are equal,
## while the alternative hypothsis is that they are not equal. According to the paired
## t-test result, we get a p-value of 0.03383 under a confidence level of 95%.
## Therefore, we reject the null hypothesis with 95% confidence and can reasonably
## conclude that means between the two groups are not equal.

t.test(before_data$trestbps, after_data$trestbps, mu = 0, alternative = 'two.sided', paired = TRUE, con
```

```
##
##  Paired t-test
##
## data:  before_data$trestbps and after_data$trestbps
## t = 2.1422, df = 146, p-value = 0.03383
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.3255737 8.0825895
## sample estimates:
## mean of the differences
##                4.204082
```

Parametric Statistical Test (One-way ANOVA)

```r
# In this example, I will conduct a One-way Analysis of Variance test (ANOVA) to
## test if the 'thalach' means for the four 'cp' groups are the same. Therefore, the
## null hypothesis is that the means for the four groups are the same, while the
```

```r
## alternative hypothesis is that the means for the four groups are different.

## Please note that the example is not strict with the assumptions and conditions
## required by the test. The sole purpose of the illustration is to show how to conduct
## the test in terms of coding. Assume that they the data set(s) meet all the necessary
## assumptions and requirements of the test to be conducted.

# First of all, plot a box plot to visualize the distribution.

ggplot(data = data_clean, aes(x = cp, y = thalach, fill = cp)) +
  geom_boxplot(show.legend = FALSE) +
  geom_point(show.legend = FALSE) +
  labs(title = 'Box plot showing the distribution of thalach for each cp group') +
  theme(plot.title = element_text(hjust = 0.5)) +
  theme_hc()
```




```r
ANOVA1 = aov(data_clean$thalach ~ data_clean$cp)
ANOVA1
```

```
## Call:
##    aov(formula = data_clean$thalach ~ data_clean$cp)
##
## Terms:
##                 data_clean$cp Residuals
## Sum of Squares       24029.83 134413.39
```

```
## Deg. of Freedom              3       299
##
## Residual standard error: 21.20243
## Estimated effects may be unbalanced
```

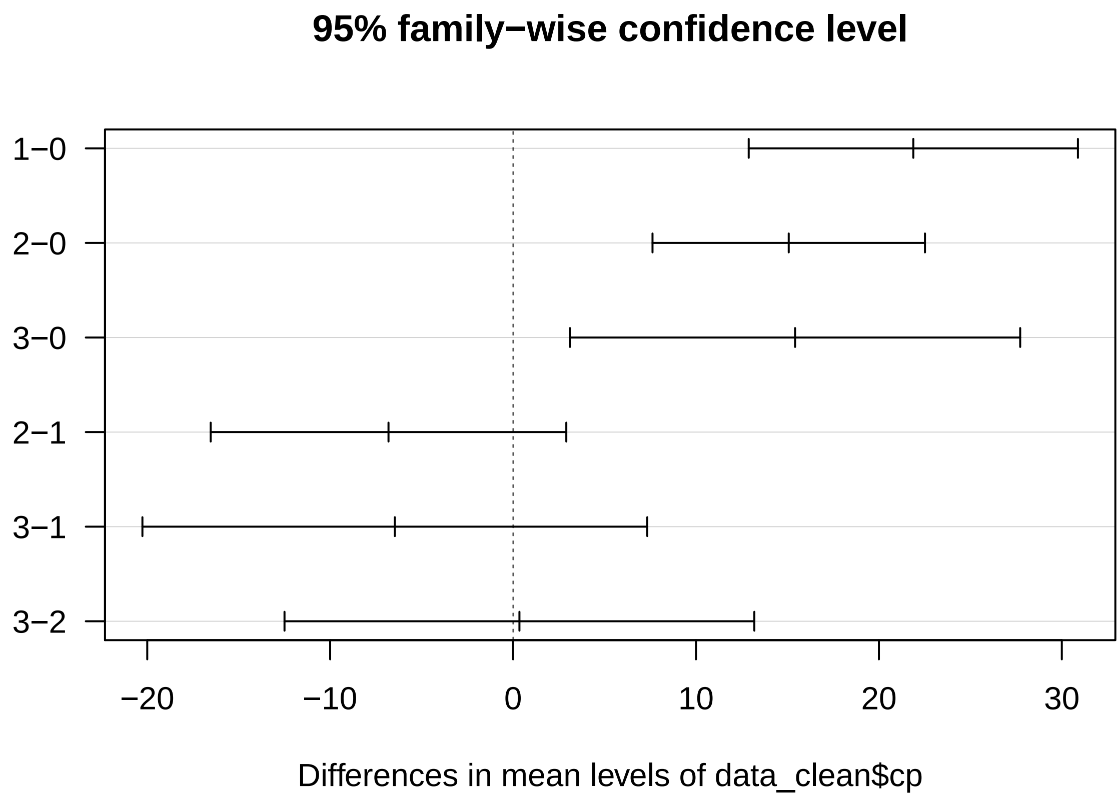
```r
# Based on both the box plot and the test result, we have evidence to conclude that
## the mean of 'thalach' of the four 'cp' groups are not equal.
summary(ANOVA1)
```

```
##                  Df Sum Sq Mean Sq F value   Pr(>F)
## data_clean$cp   3  24030    8010   17.82 1.15e-10 ***
## Residuals      299 134413     450
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# Specifically, to see which pair of 'cp' groups is different, we can use the TukeyHSD
## function.
TukeyHSD(ANOVA1)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = data_clean$thalach ~ data_clean$cp)
##
## $`data_clean$cp`
##           diff        lwr       upr     p adj
## 1-0 21.8815385  12.881882 30.881195 0.0000000
## 2-0 15.0707339   7.622787 22.518681 0.0000019
## 3-0 15.4180602   3.111904 27.724216 0.0073153
## 2-1 -6.8108046 -16.531917  2.910308 0.2705765
## 3-1 -6.4634783 -20.264550  7.337594 0.6210076
## 3-2  0.3473263 -12.495870 13.190522 0.9998774
```

```r
# Since the first three groups have a low p-value and the confidence interval does
## not include 0, we have evidence to conclude that the mean 'thalach' of 0 is
## different from that of 1, 2, and 3 cp groups.

plot(TukeyHSD(ANOVA1), las = 1)
```



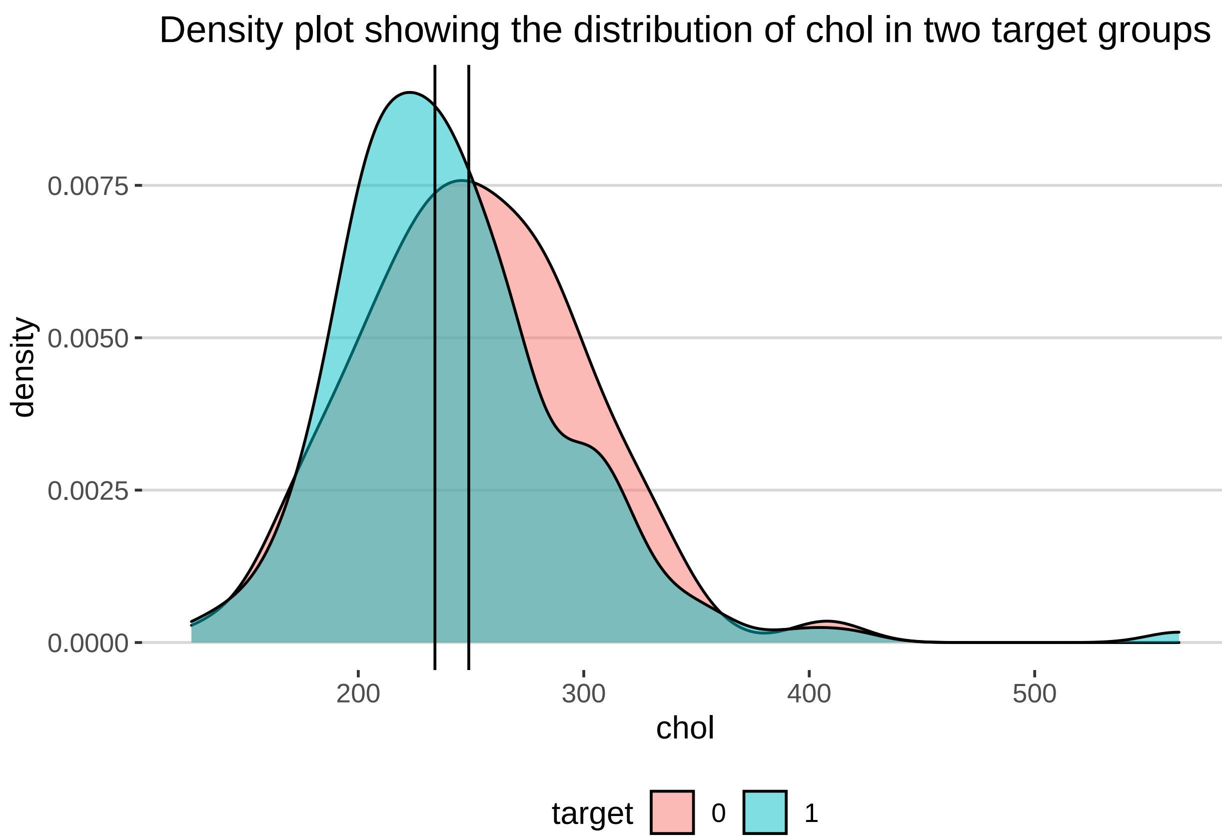
———— Nonparametric Statistical Test (Kruskal Wallis Test) ————

```
# In this example, I will conduct a Kruskal Wallis Test to test if the 'thalach'
## means for the four 'cp' groups are the same. Therefore, the null hypothesis is
## that the means for the four groups are the same, while the alternative hypothesis
## is that the means for the four groups are different.

## Please note that the example is not strict with the assumptions and conditions
## required by the test. The sole purpose of the illustration is to show how to conduct
## the test in terms of coding. Assume that they the data set(s) meet all the necessary
## assumptions and requirements of the test to be conducted.

## Based on the test result, we can see a small p-value and therefore reject the null
## hypothesis that the mean 'thalach' for the four 'cp' groups are the same.

kruskal1 = kruskal.test(data_clean$thalach ~ data_clean$cp)
kruskal1
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  data_clean$thalach by data_clean$cp
## Kruskal-Wallis chi-squared = 48.216, df = 3, p-value = 1.916e-10
```

———— Nonparametric Statistical Test (Wilcoxon Rank-Sum Test) ————

```r
# In this example, I will use the Wilcoxon Rank-Sum Test to test if the median
## 'chol' of the two target groups is different. Since Wilcoxon Rank-Sum does
## not assume known distribution, the equality of the median tested can be used
## to compare the distribution of the two groups we are comparing. For example,
## if we reject the null hypothesis that says the medians are equal, we know that
## the distribution of one group is shifted either to the left or right, thereby
## different means. Based on both the chart and the test, we can reject the null
## hypothesis that the medians of the two datasets are the same, with 95% confidence.

## Please note that the example is not strict with the assumptions and conditions
## required by the test. The sole purpose of the illustration is to show how to conduct
## the test in terms of coding. Assume that they the data set(s) meet all the necessary
## assumptions and requirements of the test to be conducted.

ggplot(data = data_clean, aes(x = chol, fill = target)) +
  geom_density(alpha = 0.5) +
  geom_vline(xintercept = c(median(data_clean$chol[data_clean$target == 0]),
                            median(data_clean$chol[data_clean$target == 1]))) +
  labs(title = 'Density plot showing the distribution of chol in two target groups') +
  theme(plot.title = element_text(hjust = 0.5)) +
  theme_hc()
```




```r
wilcox.test(data_clean$chol ~ data_clean$target, mu = 0, alternative = 'two.sided',
            conf.int = TRUE, conf.level = 0.95, paired = FALSE, correct = TRUE)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  data_clean$chol by data_clean$target
## W = 12980, p-value = 0.03572
## alternative hypothesis: true location shift is not equal to 0
## 95 percent confidence interval:
##   0.9999428 22.9999605
## sample estimates:
## difference in location
##               11.99999
```

—————— Nonparametric Statistical Test (Wilcoxon Signed Rank Test) ——————-

```r
# In this example, I will conduct a Wilcoxon Signed Rank Test to compare the median
## difference, or distribution difference, of two population that are paired. Therefore,
## the null hypothesis is that the medians of the two data sets are the same, while
## the alternative hypothesis is that they are different. According to the test result,
## we get a p-value of 0.08047. Therefore, we fail to reject the null hypothesis with
## 95% confidence level and reasonably conclude that the medians of the two groups are
## not significantly different.

## Please note that the example is not strict with the assumptions and conditions
## required by the test. The sole purpose of the illustration is to show how to conduct
## the test in terms of coding. Assume that they the data set(s) meet all the necessary
## assumptions and requirements of the test to be conducted.

wilcox.test(before_data$trestbps, after_data$trestbps, mu = 0, alternative = 'two.sided', paired = TRUE
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  before_data$trestbps and after_data$trestbps
## V = 5775.5, p-value = 0.08047
## alternative hypothesis: true location shift is not equal to 0
## 95 percent confidence interval:
##  -0.4999988  7.9999972
## sample estimates:
## (pseudo)median
##       3.999952
```

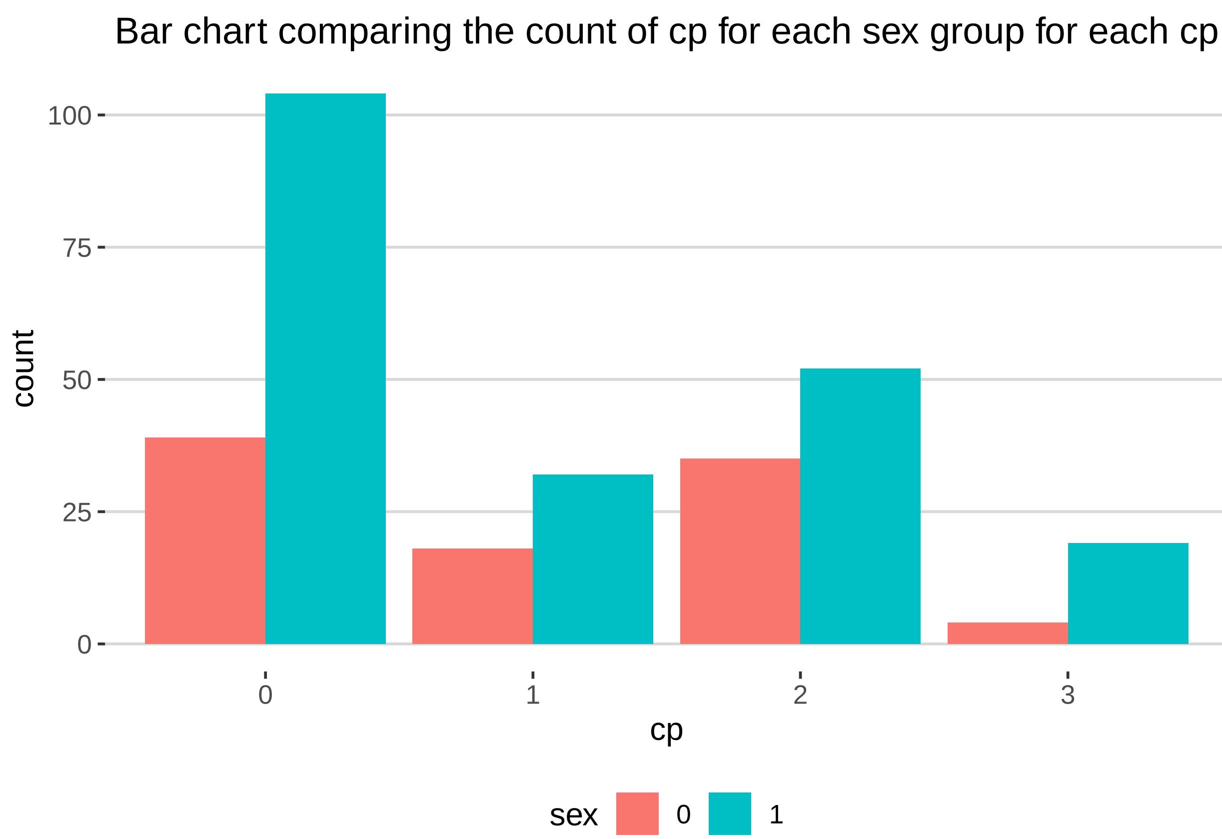
—————————— Chi-Square Test——————————

```r
# In this example, I will conduct a Chi-Square Test on two categorical variables,
## 'sex' and 'cp,' to see if the occurrence of one category is independent from
## the occurrence of another category. Therefore, the null hypothesis is that the
## two variables are independent, while the alternative hypothesis is that the two
## variables are dependent

## Please note that the example is not strict with the assumptions and conditions
## required by the test. The sole purpose of the illustration is to show how to conduct
## the test in terms of coding. Assume that they the data set(s) meet all the necessary
## assumptions and requirements of the test to be conducted.
```

```r
table1 = table(data_clean$cp, data_clean$sex)
table1
```

```
##
##       0   1
##   0  39 104
##   1  18  32
##   2  35  52
##   3   4  19
```

```r
ggplot(data = data_clean, aes(x = cp)) +
  geom_bar(data = data_clean, stat = 'count', aes(fill = sex), position = 'dodge') +
  labs(title = 'Bar chart comparing the count of cp for each sex group for each cp') +
  theme(plot.title = element_text(hjust = 0.5)) +
  theme_hc()
```




```r
# Based on both the graph and the test result, we fail to reject the null hypothesis
## and conclude that the 'sex' and 'cp' variables are dependent.
chisq.test(table1, correct = TRUE)
```

```
##
##  Pearson's Chi-squared test
##
## data:  table1
## X-squared = 6.8221, df = 3, p-value = 0.07779
```