ASSIGNMENT 11

GSI Intro to Big Data and Data Mining

The University of Texas at Austin

Zhaowen Fan
Rafael Ignacio Gonzalez Chong

Table of Contents

Task 1: Summarize the data by each feature. Use an appropriately labelled table to show the results. Also include a graphical presentation that shows the distribution of Cholesterol for participants vs. non-participants. Describe the shape of each distribution. Use R to calculate the quantities and generate the visual summaries.

```
      age              sex               cp            trestbps          chol            fbs        
 Min.   :29.00   Min.   :0.0000   Min.   :0.0000   Min.   : 94.0   Min.   :126   Min.   :0.0000  
 1st Qu.:48.00   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:120.0   1st Qu.:211   1st Qu.:0.0000  
 Median :56.00   Median :1.0000   Median :1.0000   Median :130.0   Median :240   Median :0.0000  
 Mean   :54.43   Mean   :0.6956   Mean   :0.9424   Mean   :131.6   Mean   :246   Mean   :0.1493  
 3rd Qu.:61.00   3rd Qu.:1.0000   3rd Qu.:2.0000   3rd Qu.:140.0   3rd Qu.:275   3rd Qu.:0.0000  
 Max.   :77.00   Max.   :1.0000   Max.   :3.0000   Max.   :200.0   Max.   :564   Max.   :1.0000  
    restecg          thalach         exang           oldpeak         slope             ca        
 Min.   :0.0000   Min.   : 71.0   Min.   :0.0000   Min.   :0.000   Min.   :0.000   Min.   :0.0000  
 1st Qu.:0.0000   1st Qu.:132.0   1st Qu.:0.0000   1st Qu.:0.000   1st Qu.:1.000   1st Qu.:0.0000  
 Median :1.0000   Median :152.0   Median :0.0000   Median :0.800   Median :1.000   Median :0.0000  
 Mean   :0.5298   Mean   :149.1   Mean   :0.3366   Mean   :1.072   Mean   :1.385   Mean   :0.7541  
 3rd Qu.:1.0000   3rd Qu.:166.0   3rd Qu.:1.0000   3rd Qu.:1.800   3rd Qu.:2.000   3rd Qu.:1.0000  
 Max.   :2.0000   Max.   :202.0   Max.   :1.0000   Max.   :6.200   Max.   :2.000   Max.   :4.0000  
      thal           target       
 Min.   :0.000   Min.   :0.0000  
 1st Qu.:2.000   1st Qu.:0.0000  
 Median :2.000   Median :1.0000  
 Mean   :2.324   Mean   :0.5132  
 3rd Qu.:3.000   3rd Qu.:1.0000  
 Max.   :3.000   Max.   :1.0000  
```

Fig. 1 Summary of each Feature.

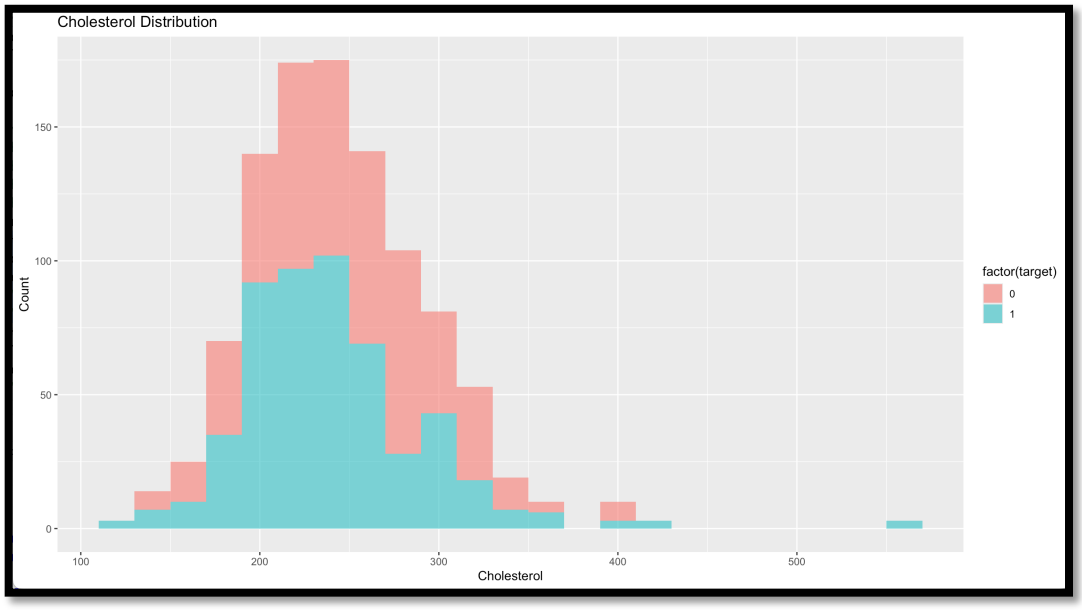

Fig. 2 Histogram of Cholesterol Distribution .

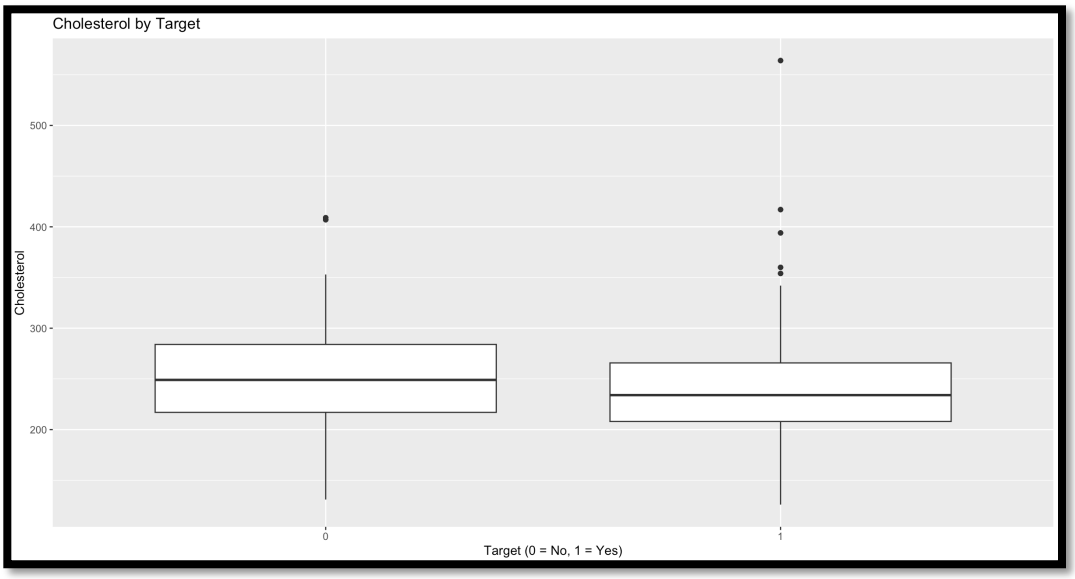

Fig. 3 Histogram of Cholesterol by Target .

Task 2: Does the mean cholesterol level is less than 250? Formally test at the alpha = 0.05 level using the 5 steps outlined in the last lecture.

```
        One Sample t-test

data:  heart$chol
t = -2.4822, df = 1024, p-value = 0.006609
alternative hypothesis: true mean is less than 250
95 percent confidence interval:
    -Inf 248.653
sample estimates:
mean of x
      246
```

Fig. 4 Task 2 result.

Task 3: Calculate a 90% confidence interval for the mean cholesterol. Interpret the confidence interval.

```
        One Sample t-test

data:  heart$chol
t = 152.65, df = 1024, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
90 percent confidence interval:
 243.347 248.653
sample estimates:
mean of x
      246
```

Fig. 5 Task 3 result.

Task 4: Formally test that resting blood pressure level is less than 130 at the alpha = 0.05 level using the 5 steps outlined in our last class.

```
        One Sample t-test

data:  heart$trestbps
t = 2.9457, df = 1024, p-value = 0.9984
alternative hypothesis: true mean is less than 130
95 percent confidence interval:
     -Inf 132.5125
sample estimates:
mean of x
 131.6117
```

Fig. 6 Task 4 result.

Task 5: Calculate a 95% confidence interval for the resting blood pressure. Interpret the confidence interval.

```
        One Sample t-test

data:  heart$trestbps
t = 240.55, df = 1024, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 130.5381 132.6853
sample estimates:
mean of x
 131.6117
```

Fig. 7 Task 5 result.

Task 6: Are the cholesterol level of the two groups with target 1 or 0 different? (Is it bigger, less or equal?)

```
        Welch Two Sample t-test

data:  heart$chol by heart$target
t = 3.2191, df = 1022.8, p-value = 0.001326
alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
95 percent confidence interval:
  4.026703 16.600292
sample estimates:
mean in group 0 mean in group 1
       251.2926        240.9791
```

Fig. 8 Task 6 result.

Task 7: Are resting blood pressure level of the two groups with target 1 or 0 different? (Is it bigger, less or equal?)

```
        Welch Two Sample t-test

data:  heart$trestbps by heart$target
t = 4.4652, df = 986.06, p-value = 8.922e-06
alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
95 percent confidence interval:
 2.724668 6.997262
sample estimates:
mean in group 0 mean in group 1
       134.1062        129.2452
```

Fig. 9 Task 7 result.

Task 8: Are the fasting blood sugar level of the two groups with target 1 or 0 different? (Is it bigger, less or equal?)

```
        Welch Two Sample t-test

data:  heart$fbs by heart$target
t = 1.3149, df = 1005.1, p-value = 0.1888
alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
95 percent confidence interval:
 -0.01445025  0.07314559
sample estimates:
mean in group 0 mean in group 1
      0.1643287       0.1349810
```

Fig. 9 Task 8 result.

Task 9: Are the maximum heart rate level of the two groups with target 1 or 0 different? (Is it bigger, less or equal?)

```
        Welch Two Sample t-test

data:  heart$thalach by heart$target
t = -14.862, df = 976.86, p-value < 2.2e-16
alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
95 percent confidence interval:
 -22.02427 -16.88631
sample estimates:
mean in group 0 mean in group 1
       139.1303        158.5856
```

Fig. 10 Task 9 result.

Appendices (Code)

```
#ASSIGNMENT 11

#GSI Intro to Big Data and Data Mining

#Zhaowen Fan

#Rafael Ignacio Gonzalez Chong


library(readr)

library(ggplot2)


heart <- read_csv("heart.csv")


#Task 1: Summarize the data by each feature. Use an appropriately labelled

#table to show the results. Also include a graphical presentation that shows

#the distribution of Cholesterol  for participants vs. non-participants.

#Describe the shape of each distribution. Use R to calculate the quantities and
```

```
#generate the visual summaries.  (2 points)
summary(heart)
ggplot(heart, aes(x = chol, fill = factor(target))) +
  geom_histogram(binwidth = 20, alpha = 0.6) +
  labs(title = "Cholesterol Distribution",
       x = "Cholesterol",
       y = "Count")
ggplot(heart, aes(x = factor(target), y = chol)) +
  geom_boxplot() +
  labs(title = "Cholesterol by Target",
       x = "Target (0 = No, 1 = Yes)",
       y = "Cholesterol")



#Task 2: Does the mean cholestoral level is less than 250? Formally test at the
#alpha = 0.05 level using the 5 steps outlined in the last lecture. (6 points)
# Hypotheses:
# H0: mean = 250
# H1: mean < 250
t.test(heart$chol, mu = 250, alternative = "less", conf.level = 0.95)



#Task 3: Calculate a 90% confidence interval for the mean cholestoral.
```

```
#Interpret the confidence interval. (4 points)
t.test(heart$chol, conf.level = 0.90)



#Task 4: Formally test that resting blood pressure level is less than 130 at
#the alpha = 0.05 level using the 5 steps outlined in our last class. (6 points)
# Hypotheses:
# H0: mean trestbps = 130
# H1: mean trestbps < 130
t.test(heart$trestbps, mu = 130, alternative = "less", conf.level = 0.95)



#Task 5: Calculate a 95% confidence interval for the resting blood pressure.
#Interpret the confidence interval. (4 points)
t.test(heart$trestbps, conf.level = 0.95)



#Task 6: Are the cholesterol level of the two groups with target 1 or 0
#different? (Is it bigger, less or equal?)
# Hypotheses:
# H0: The mean cholesterol is the same for both groups
# H1: The means are different
t.test(heart$chol ~ heart$target, alternative = "two.sided", conf.level = 0.95)
```

```
#Task 7: Are resting blood pressure level of the two groups with target 1 or 0
#different?  (Is it bigger, less or equal?)
# Hypotheses:
# H0: The mean resting blood pressure is the same for both groups
# H1: The means are different
t.test(heart$trestbps ~ heart$target, alternative = "two.sided", conf.level = 0.95)

#Task 8: Are the fasting blood sugar level of the two groups with target 1 or 0
#different?  (Is it bigger, less or equal?)
# Hypotheses:
# H0: The mean fasting blood sugar is the same for both groups
# H1: The means are different
t.test(heart$fbs ~ heart$target, alternative = "two.sided", conf.level = 0.95)

#Task 9: Are the maximum heart rate level of the two groups with target 1 or 0
#different?  (Is it bigger, less or equal?)
# Hypotheses:
# H0: The mean maximum heart rate is the same for both groups
# H1: The means are different
t.test(heart$thalach ~ heart$target, alternative = "two.sided", conf.level = 0.95)
```