Report

Boyi Ding

March 25, 2024

1 Introduction

In this report, we analyze a simulated dataset with multiple features for two distinct groups: A and B. Our objective is to obtain the differences between these groups using statistical methods.

2 Data Simulation Process

In this simulated dataset, we assume it is a medicine experiment, where group A is a placebo group and group B is a treatment group. For the features, we have set four variables: male, female, naive and experienced. Naive and experienced are the situation of the patients to differentiate if the patients used the similar treatment before. In fact, the data of gender and the patients situation are also independent.
The number this dataset is random. And gender data is generated with normal distributions, the other one is generated with log-normal distributions since the data is probably positive skewed in this situation. And the number could be the level of the white blood cell of patients, lower means more effective.

```
# Set seed to create a simulated dataset
set.seed(1)
# Number of observations for each group
nA <- 100
nB <- 100
# Create simulated data for group A
group_A <- data.frame(
  Group = rep("A", nA),
  male = rnorm(nA, mean = 15, sd = 2),
  female = rnorm(nA, mean = 20, sd = 3),
  naive = rlnorm(nA, mean = 2, sd = 0.6),
  experienced = rlnorm(nA, mean = 3, sd = 0.6)
)
# Generate simulated data for group B
```

```
group_B <- data.frame(
  Group = rep("B", nB),
  male = rnorm(nB, mean = 12, sd = 2),
  female = rnorm(nB, mean = 19, sd = 3),
  naive = rlnorm(nB, mean = 1.5, sd = 0.6),
  experienced = rlnorm(nB, mean = 2.9, sd = 0.6)
)
# Combine data for both groups
simulated_data <- rbind(group_A, group_B)
```

3 Analysis

Then we use t-tests for each feature to compare the means between groups A and B.

```
t_test_results <- lapply(simulated_data[, -1], function(x){
  t_test <- t.test(x ~ Group, data = simulated_data)
  return(t_test)
})
#  p-values from the t-tests
p_values <- sapply(t_test_results, function(x) x$p.value)
# Print the p-values
print(p_values)
```

4 Result

We have the result:

```
> print(p_values)
        male         female          navie   experienced
5.598217e-26  2.610404e-02  6.800631e-10  2.817150e-03
```

It is clearly that the two groups have the significant difference in these features.
Then We create the boxplots to visualize the distribution of each feature for groups A and B(see figure 1):

```
# Create boxplots for each feature
par(mfrow = c(1, 4))
for (feature in names(simulated_data)[-1]) {
  boxplot(simulated_data[[feature]] ~ simulated_data$Group,
          xlab = "Group", ylab = feature,
          main = paste("Boxplot of", feature),
          col = c("red", "green"))
  legend("topright", legend = c("A", "B"), fill = c("red", "green"))
```

}

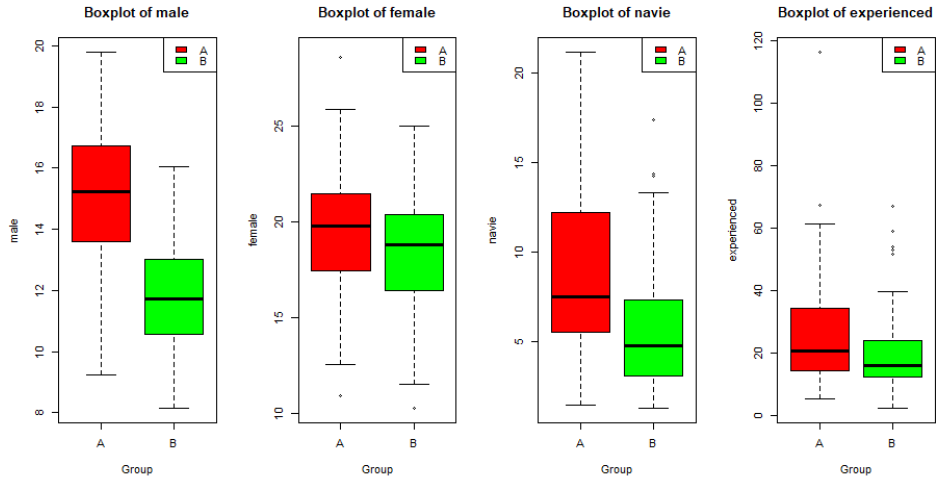

Figure 1:

It is also clear that the treatment is more useful to males. And the treatment has less effective in experienced patients which is common.