STAT 6210 Homework 2 Report
Xue Ming Wang (Vivian)
G20580112

Part A

The dataset hw2a.csv has two variables, x and y. In order to find whether X and Y have different means, the relation between the two variables needs to be determined in two situations: if X and Y are not paired, and if X and Y are paired.

1. X and Y are not paired.

1.1 Summary statistics of variable X

Moments for Variable X		Moments for Variable Y	
Mean	27.77	Mean	212.23
Variance	15,271.1668	Variance	39,316.27
Standard Error	39.07	Standard Error	62.7
CV	445	CV	93.42
Skewness	-0.55	Skewness	-0.354
Kurtosis	−0.648	Kurtosis	-1.115

Table 1. Basic statistical summary table of variable X and Y

Summary statistics in Table 1 show that there is a big difference in means, where the mean of X is 27.77 and the mean of Y is 212.2. The variance for variable Y is higher than that for variable X, which means Y is more spread out than X. The standard error for X is 39.07 and for Y is 62.7. The CV for X is 445, indicating that X has high relative variability, while the CV for Y is 93.42, indicating that Y has lower relative variability. The skewness for X is -0.55 and for Y is -0.354, both indicating a slight left skew. The kurtosis for X is -0.648, indicating a slightly flat distribution, while the kurtosis for Y is -1.115, indicating a relatively flat distribution.

1.2 Distribution of variable X and Y

Variables	Minimum	Q1	Median	Q3	Maximum
X	-191.30	-61.60	51.30	134.60	192.10
Y	-120.30	34.30	251.65	355.70	466.80

Table 2. Quantiles of variable X and Y

In table 2, The quantiles of variables X and Y depict that the range of X is wider than that of Y, as the minimum value for X is lower than that of Y and the maximum value for X is higher. This suggests that X has more extreme values compared to Y. The median for Y is much higher than that of X, indicating that the center of Y's distribution is located farther to the right than X.

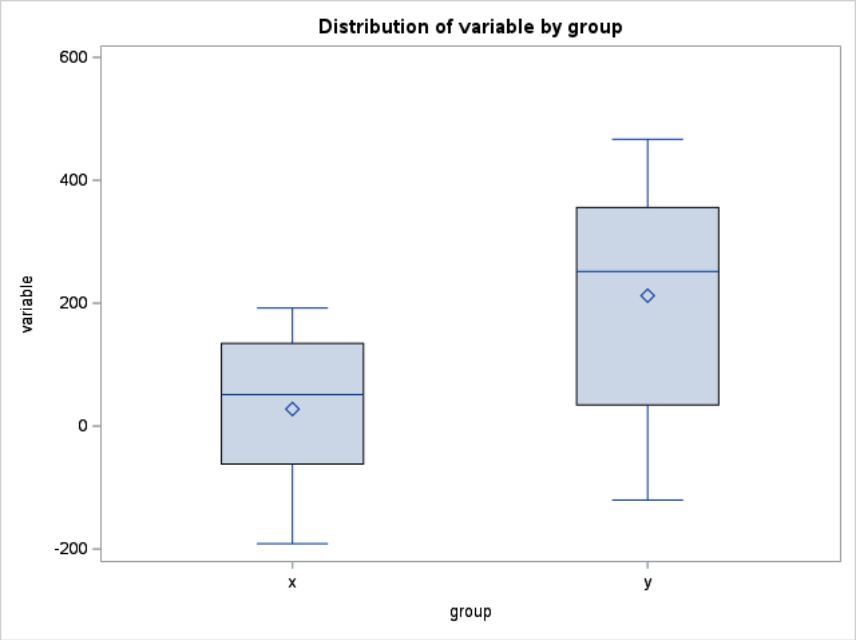

Figure 1. Boxplot for unpaired X and Y

Figure 1 displays the boxplots for both X and Y. The boxplots reveal notable discrepancies between the quantiles for X and Y. There is a significant difference between the means and medians in X and Y.

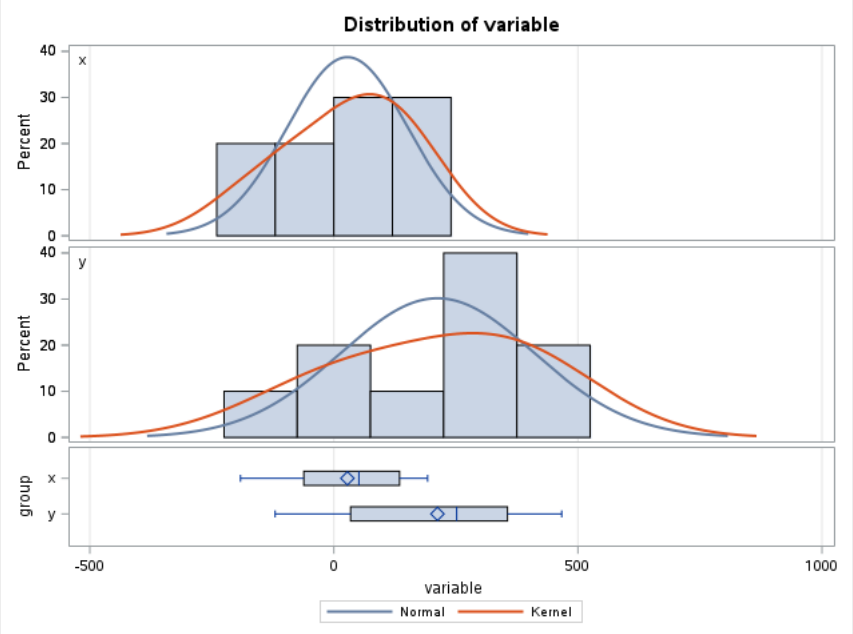

Figure 2. Distribution of X and Y

In figure 2, the distribution of the X and Y graphs showed that both variables are slightly left-skewed. Yet, it is obvious that they have different distributions. To confirm the assumption, t-tests will be used.

1.3 T-Test for unpaired X and Y

Methods	Statistic	t Value	p > |t|
Pooled	Equal	-2.5	0.0225
Satterthwaite	Unequal	-2.5	0.0246

Table 3. T-test for unpaired X and Y

Table 3 shows the results of two different t-tests conducted to compare the means of two variables. For all two tests, the t statistic is -2.5. The first test used the pooled method, which assumes that the variances of the two variables are equal. The p-value for this test is 0.0225, which is less than the significance level of 0.05. It indicated that the null hypothesis could be rejected, which means the means of the two variables are significantly different. The second test is the Satterthwaite method, which does not assume equal variances. The p-value for this test is 0.0246, which is also less than 0.05, supporting the conclusion that the means of the two variables are significantly different.

While t-tests assume certain parameters about the distribution of the variable, there are situations where we may need to assume the distribution of the variable. In such cases, nonparametric tests for location can be used. With the Wilcoxon two-sample test in table 4, the sum of scores for group x is 78, which is lower than the expected value of 105 under the null hypothesis of no difference between the two samples. The test statistic for both groups is 78, and the p-value obtained from this test is 0.0226. This indicates that there is sufficient evidence to reject the null hypothesis and conclude that there is a significant difference between the two groups. Above all, X and Y have different means when X and Y are not paired.

Group	Sum of Scores	Expected Under H0	Statistic	p value
x	78	105	78	0.0226
y	132	105	78	0.0226

Table 4. Wilcoxon Two-Sample Test for X and Y

2. X and Y are paired

In section 1, we assumed that the two variables were not paired. However, it is not reasonable to only assume that X and Y are independent, where the correlation between X and Y also needs to be considered.

2.1 Coefficient between X and Y

With the distribution of the differences between two paired variables assumed to be normal, Pearson Correlation Coefficients can detect the difference between the two means. The Pearson Correlation Coefficient between X and Y is -0.8412, indicating a weak negative correlation between the two variables. Hence, the influence of correlation can be ignored.

2.2 T-Test for paired X and Y

Difference: X − Y	
N	20
Mean	-184.5
Standard Deviation	301.3
DF	19
t-value	-2.74
p > |t|	0.0131

Table 5. T-TEST for paired X and Y

Table 5 demonstrates the results of a t-test comparing the difference between X and Y. The t-value obtained from this test was -2.74, with 19 degrees of freedom. The p-value for this test is 0.0131, which indicates that there is sufficient evidence to reject the null hypothesis and conclude that there is a significant difference between X and Y.

3. Conclusion

Summary statistics, distribution, t-tests for both paired and unpaired X and Y, and correlation between X and Y all suggest that the difference between X and Y is statistically significant, no matter whether X and Y are paired or not. Therefore, it can be concluded that the means of the two variables, X and Y, are significantly different from each other.

Part B

Gender	Treatment	Response	
		Better	Same
Female	Active	16	11
	Placebo	5	20
Male	Active	12	16
	Placebo	7	19

The above dataset is the Migraine study, which consists of hypothetical data for a clinical trial of migraine treatment. Three types of data are included: gender, treatment, and response. Both male and female receive either an active treatment or a placebo. Each would give a response of "Better" or "Same." This part of the report will seek to understand the relationship between the treatment and the response. As the variable gender might be an influence factor, both situations, conditional and not conditional on gender, will be considered in the following report.

1. Not Conditional on Gender

Statistic	Prob
Chi-Square	0.0037
Mantel-Haenszel Chi-Square	0.0038
Fisher's Exact Test	0.0032

Table 6. Statistic for unconditional gender

First, the assumption that response and treatment do not depend on gender is made in this section. I choose to use chi-square to measure the difference between observed and expected frequencies in a contingency table. The null hypothesis is that response and treatment are independent, and they have no association between them. Mantel-Haenszel Chi-Square is also conducted, which is used for stratified contingency tables, where the association between two variables is examined while controlling for the effect of the third variable. Lastly, with the same null hypothesis, Fisher's Exact Test ensures the correctness of the p-value. Thus, these three tests are performed to determine the association between two variables. Table 6 presents the results of the three statistical tests. They indicate that all the p-values are less than 0.05, suggesting that the results are statistically significant. Therefore, the response and treatment are not independent, and they have an association between them. However, the influence of gender needs to be considered as well.

2. Conditional on gender

2.1 Odd Ratio

Gender	Odd ratio
Female	5.8182
Male	2.0357

Table 7. Odd Ratio

Without the variable of gender, the odds ratio would be 3.0009. However, table 7 shows that the odds ratio for females is 5.8182 and for males is 2.0357, which are substantially different. So, it is necessary to consider the influence of gender in the subsequent analysis.

2.2 Female

Statistic	Prob
Chi-Square	0.0039
Mantel-Haenszel Chi-Square	0.0043
Fisher's Exact Test	0.0036

Table 8. Statistic for gender=female

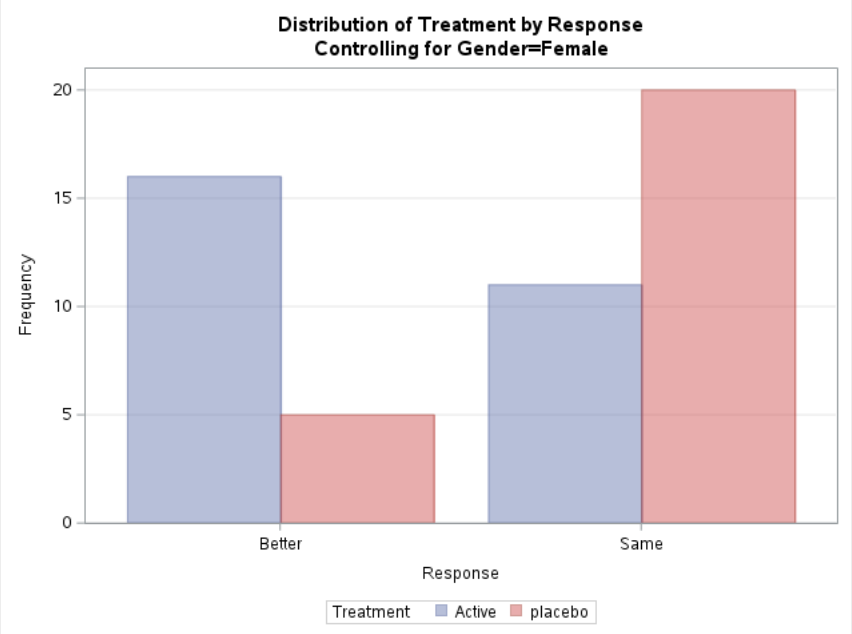

Figure 3. Distribution of treatment by response controlling for gender = female

The same three tests (table 8) that apply to unconditional on gender are now conducted under female conditions. It is apparent that there is significant evidence of a correlation between treatment and response, which is evidenced by p-value that are less than 0.05. Based on the graph in figure 3, which illustrates the distribution of response to treatment among females, the gap between the frequency of treatment and response is large. The frequency of "better" with active treatment is greater than placebo. At the same time, the frequency of "same" with the placebo is

greater than that of active treatment. It means there is a significant difference in treatment and response, and active treatment is better than the placebo for females.

2.3 Male

Statistic	Prob
Chi-Square	0.2205
Mantel-Haenszel Chi-Square	0.2249
Fisher's Exact Test	0.1090

Table 9. Statistic for gender=male

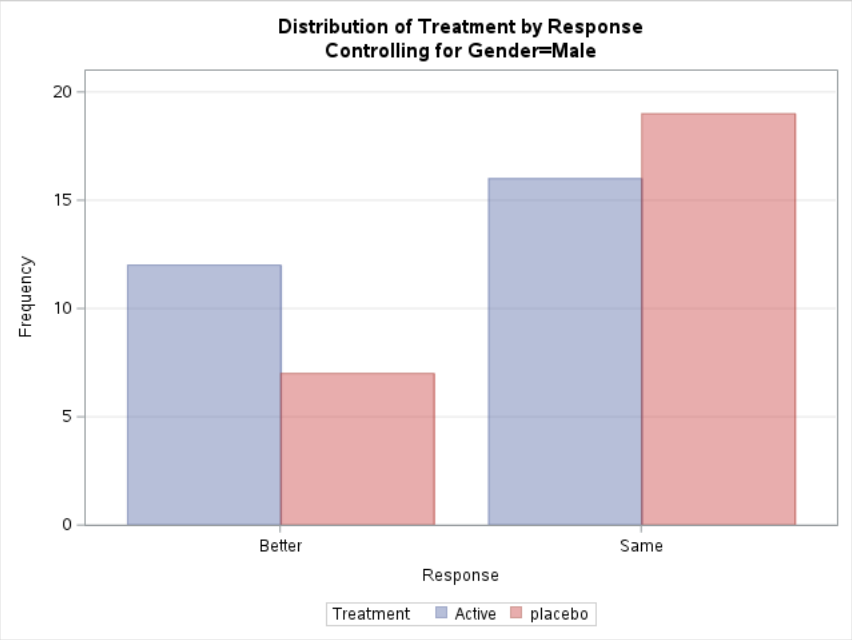

Figure 4. Distribution of treatment by response controlling for gender = male

Same as female, the three tests (table 9) are conducting under male as well. It is apparent that there is significant evidence that there is no correlation between treatment and response, which is evidenced by a p-value that is greater than 0.05. Figure 4 illustrates the distribution of responses to treatment among males. The frequency of "better" with active treatment is still greater than placebo, but the gap between them is smaller. At the same time, the frequency of "same" with placebo and active treatment is about the same. It means there is not much difference between active treatment and placebo for males.

2.4 Mantel-Haenszel Test

Statistic	Alternative Hypothesis	DF	Value	Prob
1	Nonzero Correlation	1	8.3052	0.0040
2	Row Mean Scores Differ	1	8.3052	0.0040
3	General Association	1	8.3052	0.0040

Table 10. Cochran-Mantel-Haenszel Statistics

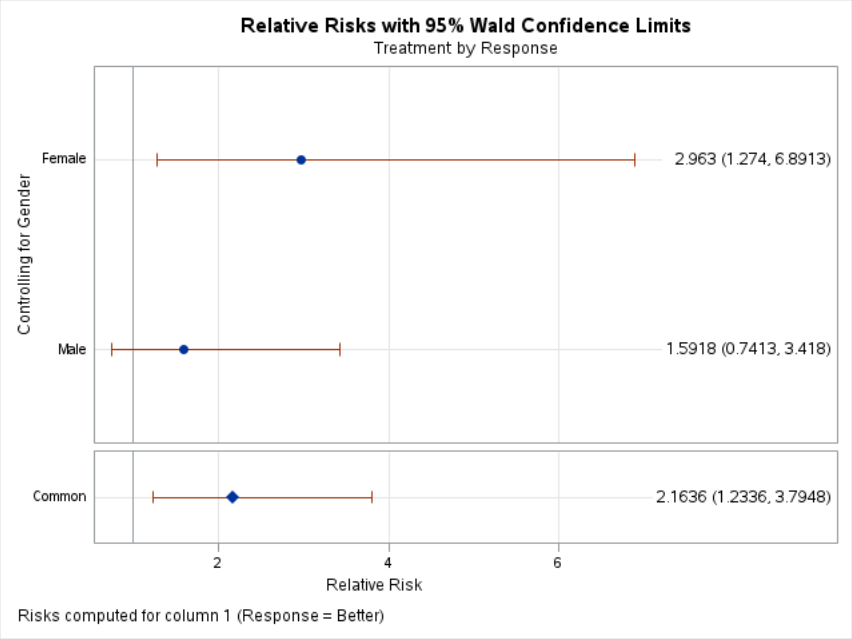

Figure 5. Distribution of treatment by response controlling for gender = male

The presented table and figure depict the findings of the Mantel-Haenszel test. With the Mantel-Haenszel test, the association can be tested. Under null, there is an association between treatment and response. In table 10, the three tests have the same hypothesis. The results reveal that the p-value remains significant at 0.004 with a degree of freedom of 1, even after adjusting for gender, which means that the relationship between treatment and response is still robust. Figure 5 illustrates the relative risks and confidence intervals for both gender levels, as well as the overall relative risk. As a result, it can be concluded that active treatment does provide better responses when considering gender.

3. Conclusion

Two situations are taken into consideration to see whether active treatment can provide better responses. If response and treatment do not depend on gender, active treatments have a strong relationship with better responses. When response and treatment depend on gender, active treatment does provide better responses for females, but there is no difference for males. Overall, by considering gender, active treatment provides better responses for females.