Homework 2

Jiaying Liu
G45268292

February 21, 2018

1 Part A

The dataset hw2a.csv contains two variables x and y. We want to explore the relation between them. And we find there are two kinds of situation such that if x and y are not paired or if they are paired. So in the following step, we will discuss the conclusion in these two situation.

1.1 When variable are not paired

1.1.1 basic statistic

To compare two unpaired variable. firstly, let us explore the basic statistic of two variable.

Table 1. Basic statistical summary of variable x and y.

x		y	
Mean	27,77	Mean	212,23
Variance	15 271,1668	Variance	39 316,27
SE	39,07	SE	62,7
CV	445	CV	93,42
Skewness	−0,55	Skewness	−0,354
Kurtosis	−0,648	Kurtosis	−1,115

Figure 1. Boxplot of variable x and y

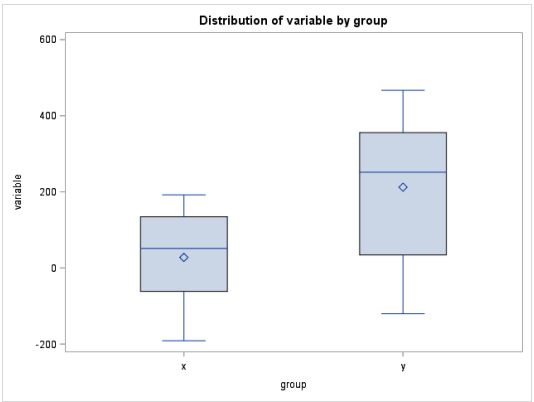


Table 1 show the basic statistical moment of two variables. Figure 1 show the boxplot of two variable. From this two results, we can find that the mean of x is 27.77 while the mean of y is 212.2 which is very different from each other. What's more, from the boxplot figure, we can find there is a big gap of each quantile between different variable. In order to further confirm the assumption, we use t-test.

1.1.2 T-Test

Table 2. T-test for x and y.

Method	Variances	t value	p $> \|t\|$
Pooled	Equal	−2,5	0,0225
Satterth-waite	Unequal	−2,5	0,0246
Cochran	Unequal	−2,5	0,0340

Figure 2. distribution x and y

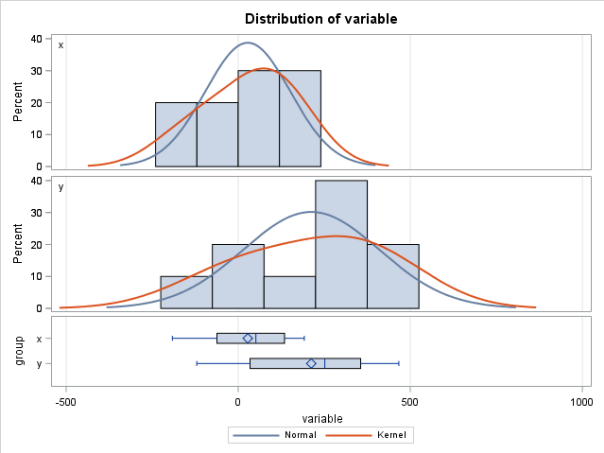


Table 2 show the t-test result and figure 2 visualize the location result. From above we can know that two variable have different variance so that we should not use the pooled method but satterthwaite and cochran method. This two method is used under the situation that variances are unequal. We can see that their p value both smaller than 0.05 which mean reject the original hypothesis and accept the alternative hypothesis. In conclusion, x and y have different mean. From the figure 2, it is very obvious that they have different distribution. However t-test is under parameter, but if we do not know the parameter of variable or we do not need the variable follows any distribution.

1.1.3 Nonparametric tests

In this section, we use NPAR1WAY procedure performs nonparametric tests for location. By the SAS result, the Wilcoxon two-sample test statistic equals 78, which is the sum of the

Wilcoxon scores for the x. This sum is smaller than 105, which is the expected value under the null hypothesis of no difference between the two samples, x and y. The one-sided p-value is 0.0226, which indicates that the x is significantly smaller than y.

Table 3. NPAR1WAY for x and y.

group	sum of scores	expected under H0	statistic	p value
x	78	105	78	0,0226
y	132	105	78	0,0226

1.2 When variable are paired

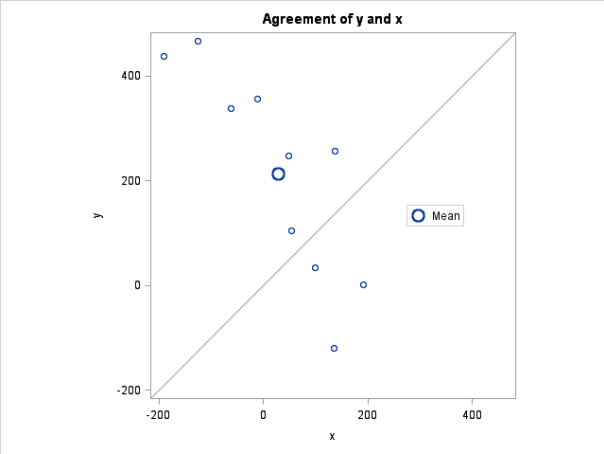
In above analysis, we assumption that two variable are not paired. However when it is not feasible to assume that two groups of data are independent, and a natural pairing of the data exists, it is advantageous to use an analysis that takes the correlation into account. Using this correlation results in higher power to detect existing differences between the means. The differences between paired observations are assumed to be normally distributed.When we test the Pearson Correlation Coefficients between two variables, we find that the value is -0.084119 which is high relative. So we can not ignore their influence of correlation.

Table 4. The TTEST procedure.

Difference: $X - Y$					
N	Mean	Std Dev	DF	t value	p $> \|t\|$
20	−184,5	301,3	19	−2,74	0,0131

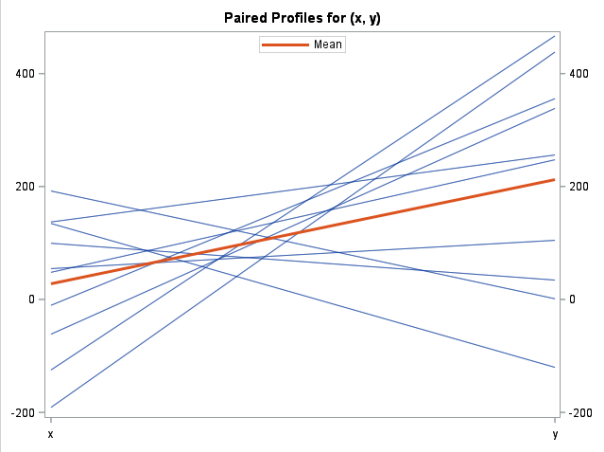
Form above table, it show the result of paired t-test. The test is used to test whether the mean change in x and y is significantly different from zero.The summary statistics of the difference are displayed,the t test is significant which $t = –2.74, p = 0.0131$, indicating that two variable have different mean.

Figure 3. Agreement of x and y



The agreement plot in figure3 reveals that only three x higher value than y.

Figure 4. Paired profiles for x and y



The profiles plot in figure 4 is a different view of the same information contained in figure 3.

2 Part B

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

The data set Migraine contains hypothetical data for a clinical trial of migraine treatment. It contain three variables which genders response and treatment. Subjects of both genders receive either a new drug therapy or a placebo. Their response to treatment is coded as 'Better' or 'Same'. We are interested in the associated between the treatment and the response. However there is a gender variable may or may not influence the result. So it will raise three question: (1) Should we combine the table neglect the influence gender; (2) Whether their odd ratio is equal or not? (3) Whether their odd ratio equal to one?

2.1 Unconditional on gender

Let's consider the situation when response and treatment do not depend on gender.

Table 5. Statistic for gender=female

statistics	prob
Chi-square	0,0037
Likelihood Ratio Chi-square	0,0033
Continuity Adj. Chi-square	0,0068
Mantel-Haenszel Chi-square	0,0038
Fisher's Exact Test	0,0032

From the table 5, displaying different statistic test.All the function of those test are same and are to verify the effect of treatment. From the result we find all the p value are smaller than 0.05 which mean the result is significant.However should we ignore the influence of gender? If we directly combine two table into one table. The odd ratio of combine table is 3.0009. But from the table 6, we find that female's odd ratio is 5.8182 and male's is 2.0357 which are so different that can not combine two table. So in the following step. We consider the influence of gender.

2.2 Stratified Table

To answer the question one, we need to test the odd ratio of different gender.

Table 6. The Odd ratio.

gender	odd ratio
female	5,8182
male	2,0357

From the above table, showing that the odd ratio of female is much large than male. It would be inappropriate to combine two table because we can not neglect the influence of gender. So we need to use stratified table to stratify gender.

2.3 Female

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

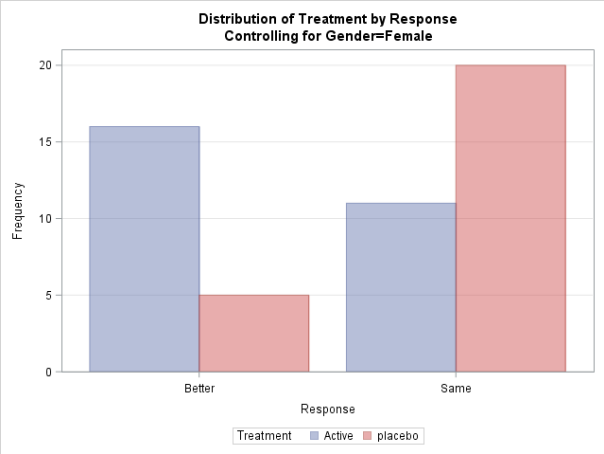


Table 7. Statistic for gender=female

statistics	prob
Chi-square	0,0039
Likelihood Ratio Chi-square	0,0033
Continuity Adj. Chi-square	0,0093
Mantel-Haenszel Chi-square	0,0043
Fisher's Exact Test	0,0036

From the figure 5, plotting the distribution of treatment by response under female.From the table 6, it show the different test for female. And we can see that there is strong evidence of association between treatment and response cause the p value is smaller than 0.05.

2.4 Male

Figure 6. Distribution of treatment by response controlling for gender = male

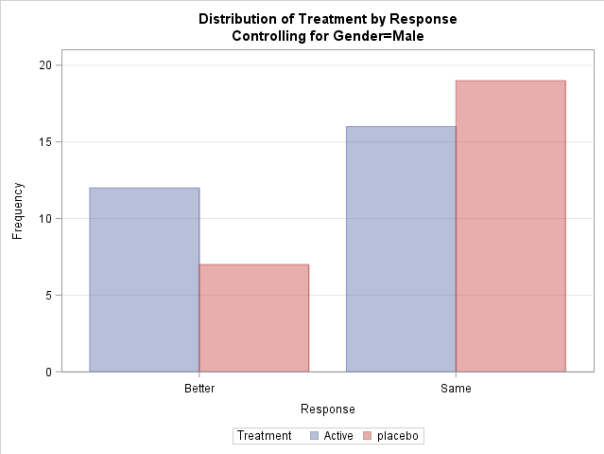


Table 8. Statistic for gender=female

statistics	prob
Chi-square	0,2205
Likelihood Ratio Chi-square	0,2184
Continuity Adj. Chi-square	0,3472
Mantel-Haenszel Chi-square	0,2249
Fisher's Exact Test	0,1090

From the figure 6, plotting the distribution of treatment by response under female.From the table 7, it show the different test for male. And we can see that there is not evidence of association between treatment and response cause the p value is higher than 0.05.

2.5 CMH Test

Now let us use CMH test to test the association.

Figure 7. Distribution of treatment by response controlling for gender = male

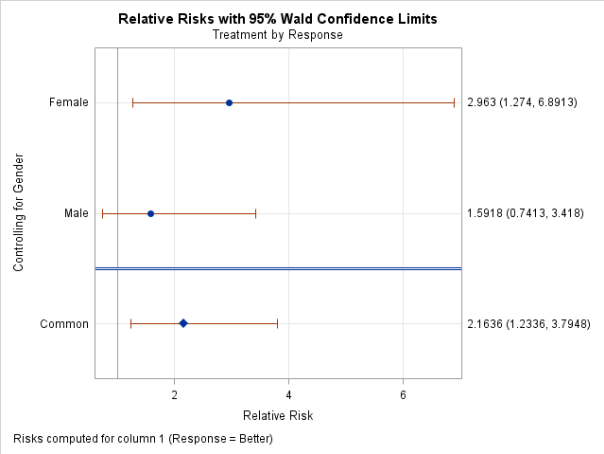


Table 9. Cochran-Mantel-Haenszel Statistic

Alternative Hypothesis	Value	prob
Nonzero Correlation	8,3052	0,004
Row Mean Scores Differ	8,3052	0,004
General Association	8,3052	0,004

Above table and figure show the results of the CMH test. figure 7 displays the relative risks and confidence limits for the two levels of Gender and for the overall relative risk. Table 8 show the CMH statistics. For a stratified table, the three CMH statistics test the same hypothesis. The significant p-value is 0.004 indicates that the association between treatment and response still strong after adjusting for gender. In conclusion, by considering the gender, Active treatment do provide better responses.