HW 2 Report

Yongchao Qiao

Part A

1. Input data into SAS manually

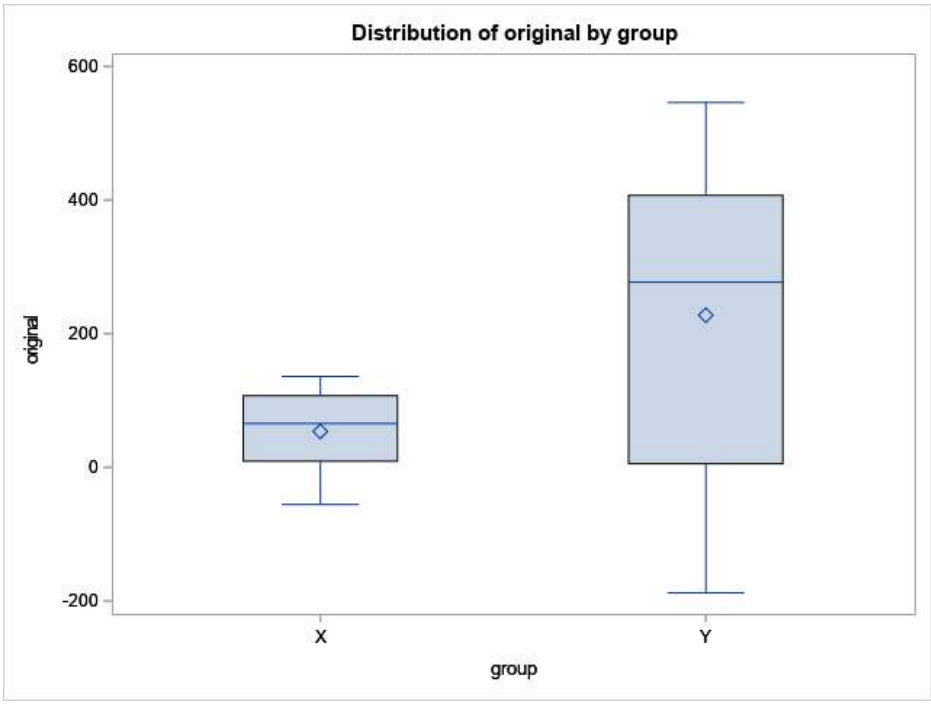
There are two methods to input the data into SAS. The first one is to use proc import procedure to read the given csv file directly. Then another one is to use the cards function to input the data by data procedure. Both these two methods have been implemented and the results are the same, saved in the datasets named HW2p and HW2d respectively.

2. Data visualization

a. Boxplots for comparing two variables

The boxplots for the two variables are shown as below. First, it indicates that these two boxes have the overlap area. Then medians will be taken into consideration. It shows that the median line of variable Y lies outsides the box of variable X entirely, then there is likely to be a difference between these two variables. Also, they have the different means, with the IQR of variable X being obviously smaller than that of variable Y. In general, one conclusion can be drawn as variable X and variable Y have different distributions.

Figure 1



b. Q-Q plots for evaluating normality (both X and Y)

The Q-Q plots for variable X and Y are shown as below. Since variable X and Y has different means and variance, there are different normal lines displayed on the plots. Both these two plots show that the dots are distributed around normal line respectively, which means both variable X and variable Y obey to the normal distribution.

Figure 2

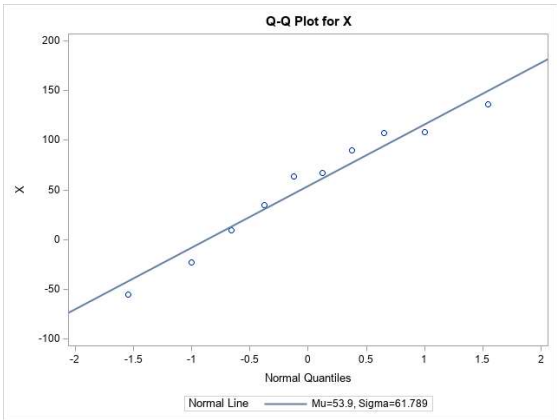
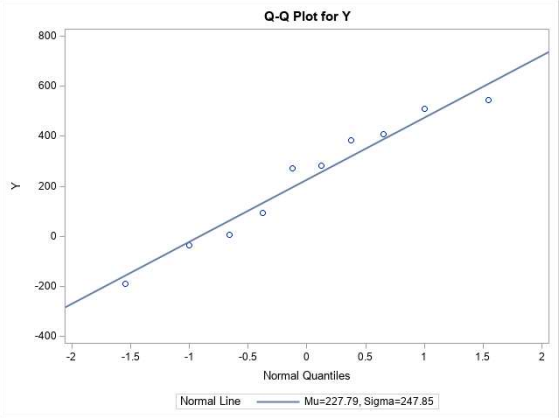


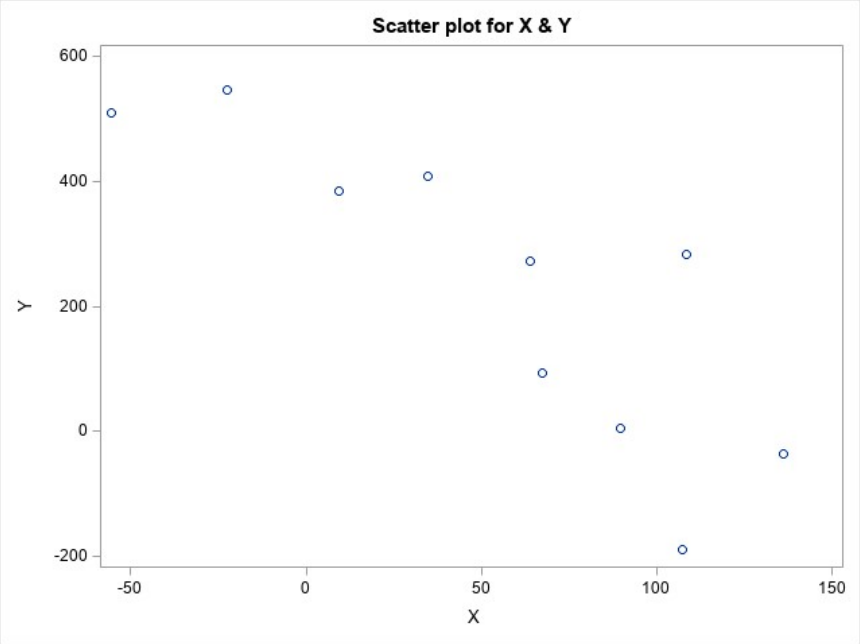
Figure 3



c. Scatter-plot for evaluating the correlation between X and Y

The scatter plot is shown as below. In general, there is an obvious linear pattern between variable X and Y. Specifically, the slope's sign of the linear pattern is negative. Therefore, it can be inferred that there is a negative correlation relationship between the variable X and variable Y.

Figure 4



3. Statistical tests

a. Test two population variances

$$H_0 : \sigma_1 = \sigma_2, \quad H_1 : \sigma_1 \neq \sigma_2$$

In this step, though it has not been verified that both variable X and variable Y follow the normal distribution with the independence between variable X and variable Y, we just assume the assumptions are met to continue the Folded F test (Although it might not be valid). The result of two population variance test is shown as table 1. Since P values of Folded F test is less than 0.05, so we can reject the null hypothesis that $\sigma_1 = \sigma_2$. That is, variable X and variable Y do not have the equal variance.

Table 1: Two population variances test

Test	Num DF	Den DF	F value	Pr > F
Folded F	9	9	16.09	0.0003

b. Perform two-sample t-test and Wilcoxon rank sum test

Two-sample t-test

$$H_0 : \mu_1 = \mu_2, \quad H_1 : \mu_1 \neq \mu_2$$

In this step, though it has not been verified that both variable X and variable Y follow the normal distribution with the independence between variable X and variable Y, we just assume the assumptions are met to continue the two-sample t-test (Although it might not be valid). Based on previous Folded F test, variable X and variable Y do not have the equal variance. Thus, the Satterthwaite method with unequal variance will be considered. Since P values of Satterthwaite method is greater than 0.05, so we cannot reject the null hypothesis that $\mu_1 = \mu_2$. That is, variable X and variable Y have the equal means.

Table 2: Two-sample t-test

Method	Variance	DF	t value	Pr > |t|
Pooled	Equal	18	-2.15	0.0451
Satterthwaite	Unequal	10.114	-2.15	0.0565

Wilcoxon rank sum test

H_0: Variable X and variable Y have the same distribution,

H_1: Variable X and variable Y do not have the same distribution.

Also, for the Wilcoxon rank sum test, we just assume variable X and variable Y meet the assumption (Although it is might not be right). Since there are 10 observations both in variable X and variable Y, less than 15, Kruskal-Wallis test is more effective than Wilcoxon rank sum

test. Although in the table 2, the two-sided P value of normal approximation is greater than 0.05, indicating that null hypothesis cannot be rejected. The P value of Kruskal-Wallis test is

Table 3: Wilcoxon rank sum test

Method	Z	One-Sided Pr < Z	Two-Sided Pr > |Z|
Normal Approximation	-1.4741	0.0702	0.0784
t Approximation	.	0.1405	0.1568

0.1306, greater than 0.05 in the table 3, so the null hypothesis that variable X and variable Y have the same distribution cannot be rejected. Then the conclusion that variable X and variable Y have the same distribution can be drawn. Although the conclusions of Kruskal-Wallis test and Wilcoxon rank sum test are the same, technically the numbers of observations in both populations refer that Kruskal-Wallis test is more appropriate.

Table 4: Kruskal-Wallis Test

Test	Chi-Square	DF	Pr > Chi-Square
Kruskal-Wallis	2.2857	1	0.1306

c. Test the correlation between X and Y

$$H_0 : \text{Rho} = 0, \quad H_1 : \text{Rho} \neq 0$$

The result is shown as table 4. The correlation coefficient between X and Y is -0.84122 with P value as 0.0023, less than 0.05. Thus, it indicates that the null hypothesis that Rho = 0 can be rejected. That is, variable X and variable Y have a significant negative correlation relationship, with Pearson correlation coefficient as -0.84122.

Table 5: Pearson Correlation Test between X and Y|

	X	Y
X	1.00000	-0.84122(0.0023)
Y	-0.84122(0.0023)	1.00000

d. Perform pared t-test and Wilcoxon signed rank test

The paired *t*-test requires the normality assumption, so a normality test with respect to the difference will be implemented first.

Normal test:

H_0 : The difference follows the normal distribution

H_1 : The difference does not follow the normal distribution

The results of normal distribution tests are shown as table 5. Since P values of Cramer-

von Mises, Kolmogorov-Smirnov and Anderson-Darling test are all greater than 0.05, so we cannot reject the null hypothesis that the difference between two populations follows the normal distribution. Specifically, Kolmogorov-Smirnov test focuses on the largest vertical

Table 6: Normal distribution tests

Test	Statistics		P-value	
Kolmogorov-Smirnov	D	0.145	Pr>D	>0.150
Cramer-von Mises	W-sq	0.030	Pr>W-sq	>0.250
Anderson-Darling	A-sq	0.218	Pr>A-sq	>0.250

difference, while Cramer-von Mises focuses on the overall squared difference with Anderson-Darling focusing on the weighted overall squared difference between the observed distribution and the theoretical normal distribution. However, no matter from which aspect, the difference between two populations follows the normal distribution.

Pared t-test and Wilcoxon signed rank test

$$H_0 : \text{Difference} = 0, \qquad H_1 : \text{Difference} \neq 0$$

In this step, the difference follows the normal distribution, so the paired t-test can be implemented. Then, for the Wilcoxon signed rank test, since the difference follows the normal distribution, the difference meets the symmetric distribution assumption. The results of these two tests are shown as table 6. Since P values of paired t-test and signed rank test are all greater than 0.05, so we cannot reject the null hypothesis that Difference = 0, for paired t-test with respect to mean while for signed rank test with respect to median.

Table 7: Pared t-test and Wilcoxon signed rank test

Test	Statistics		P-value	
Paired t	t	-1.823	Pr>|t|	0.1017
Signed Rank	S	-16.5	Pr>=|S|	0.1055

4. Conclusion

There is not a difference between X and Y populations. Basically, there is a negative correlation relationship between variable X and variable Y, so two-sample t-test will not be valid. When variable X and variable Y are not paired, Wilcoxon rank sum test, being valid, indicates that variable X and variable Y have the same distribution. When regarding variable X and variable Y as paired samples, Pared t-test and Wilcoxon signed rank test, being valid, indicate that variable X and variable Y have the same distribution. No matter paired or not, the conclusions are consistent. Therefore, there is not a difference between X and Y populations.

Part B

1. Perform a test on the association between the treatment and response

H_0 : Treatment and response are independent

H_1 : Treatment and response are not independent

Since the data is a 2x2 contingency table, Chi-Square test will be appropriate to test on the association between the treatment and response. The result is shown as table 7. The P value of Chi-Square Test is 0.0037, less than 0.05, so the null hypothesis that treatment and response are independent can be rejected. That is, there is an association between the treatment and response.

Table 8: Chi-Square Test

Statistic	Value	DF	Prob
Chi-Square	8.4429	1	0.0037

2. Test the association between treatment and response with Gender's confounding effect

Since the data is a stratified table when considering confounding effect of Gender, so Mantel-Haenszel test will be implemented to test the association between treatment and response. However, before applying the Mantel-Haenszel test, Breslow-Day Test for homogeneity of the odds ratios should be implemented first. These two tests' results are shown as below.

Breslow-Day Test

H_0 : The odds ratios are homogeneous

H_1 : The odds ratios are not homogeneous

Table 9: Breslow-Day Test for homogeneity of the odds ratios

Statistic	Value	DF	Prob
Chi-Square	1.4929	1	0.2218

The result is shown as table 8. The P value of Breslow-Day Test is 0.2218, greater than 0.05, so the null hypothesis that the odds ratios are homogeneous cannot be rejected. That is the degree and direction of the association between treatment and response are the same in each stratum. Then this conclusion validates the Mantel-Haenszel test for testing the association between treatment and response with considering the confounding effect of Gender.

Mantel-Haenszel test

H_0 : Treatment and response are independent

H_1 : Treatment and response are not independent

The result is shown as table 9. The P values of all statistics for Mantel-Haenszel Test are 0.004, less than 0.05, so the null hypothesis that treatment and response are independent can be rejected. That is, there is an association between the treatment and response with the consideration of possible confounding effect of Gender.

Table 10: Mantel-Haenszel Test

Statistics	Value	DF	Prob
Nonzero Correlation	8.3052	1	0.0040
Row Mean Scores Differ	8.3052	1	0.0040
General Association	8.3052	1	0.0040

3. Conclusion

Based on previous two parts' analysis, there is an association between the treatment and response, no matter with the consideration of possible confounding effect of Gender or not.

Appendix:

```
/*Part A*/
proc import datafile = 'E:/GW/Textbook/Data Analysis/HW2/HW2.csv' /*read the file into
sas*/
dbms = csv /*specify the format of the file*/
out=work.HW2p;      /*specify the saved dataset in sas*/
getnames=yes;      /*get the name of the variables from the original file*/
run;    /*run this procedure*/

proc print data=HW2p;    /* print the dataset saved in sas*/
run; /*run the print procedure*/

data HW2d;
input X Y;
cards;
107.3     -187.9
89.8       5.4
67.3       93.3
-55.6      510.3
136.1      -35.9
108.6      282.4
-22.5      546
64   271.8
34.8       407.1
9.2  385.4
; /* Use the cards function to read the data into HW2d*/
proc print data=HW2d;    /* print the dataset saved in sas*/
run; /*run the print procedure*/


data HW2g; /* Create the dataset HW2g */
set hw2p; /* Read the dataset HW2p */
original=Y; group='Y';output;
original=X; group='X';output; /*Reorganize the data into one variable named original with
two groups: X and Y */
drop X Y; /*Delete the original variable X and Y */
run;
proc sort data=HW2g;
by group;    /* Sort the dataset HW2g with the ascending order */
run;
proc boxplot data=HW2g;
plot original*group/boxstyle=schematic;      /* Plot the boxplot for variable original by
groups */
```

```
run;

proc univariate data=HW2d;
qqplot x / normal(mu = 53.90000 sigma = 61.78905);
qqplot y / normal(mu = 227.7900 sigma = 247.84787); /* Plot the Q-Q plot of variable X and
Y with corresponding normal lines */
run;
title 'Scatter plot for X & Y'; /* Add the plot title */
proc sgplot data=HW2p;
scatter x=X y=Y; /* Plot the scatter plot of variable X and Y */
run;
title; /*Cancel the previous title of the plot */

proc ttest data=HW2g;
var original;
class group; /* Implement the two-sample t-test with respect to two groups*/
run;

proc npar1way data=Hw2g wilcoxon;
var original;
class group; /* Implement the Wilcoxon Mann-Whitney rank sum test*/
run;
proc corr data=HW2p;
var X Y; /* Estimate and test the Pearson correlation coefficient */
run;

data hw2gp;
set hw2p;
  difference=X-Y; /* Generate the difference variable */
run;

proc ttest data=HW2gp;
   paired X*Y; /* Implement the paired t-test */
run;
proc univariate data=HW2gp;
   var difference;
   histogram difference /normal; /* Test the normality of the variable difference and
implement the one sample t-test as well as the signed rank test */
run;


/* Part B */
data Migraine;
input Gender $ Treatment $ Response $ Count @@;
```

```
datalines;
female Active    Better 16      female Active    Same 11
female Placebo Better    5      female Placebo Same 20
male      Active    Better 12      male      Active    Same 16
male      Placebo Better    7      male      Placebo Same 19
; /*Read the data into sas and save the data into dataset named Migraine */

proc freq data=Migraine order=data;
   tables Treatment*Response / chisq; /* Implement the chi-square test on the 2*2
contingency table */
   weight Count; /*Indicate the frequency based data */
run;

proc freq data=Migraine order=data;
   tables Gender*Treatment*Response / cmh; /* Implement the CMH test on the stratified
table */
   weight Count; /*Indicate the frequency based data */
run;
```