

PS4

2024-09-18

STAT 5000 HOMEWORK #4

FALL 2024 DUE FRI, SEP 27TH @ 11:59 PM NAME:

COLLABORATORS:

Directions: Type or clearly handwrite your solutions to each of the following exercises. Partial credit cannot be given unless all work is shown. You may work in groups provided that each person takes responsibility for understanding and writing out the solutions. Additionally, you must give proper credit to your collaborators by providing their names on the line below (if you worked alone, write “No Collaborators”):

1.

Consider the dataset from Homework #1 about survival times (in days) of guinea pigs that were randomly assigned either to a control group or to a treatment group that received a dose of tubercle bacilli (a bacterium that causes tuberculosis). These data are found in the `guinea_pigs.csv` file located in our course’s shared folder on SAS Studio. Suppose the researchers want to test the hypothesis that the mean (or median/distribution of) survival times are the same for controls and the guinea pigs infected with tubercle bacilli against the one-sided alternative that infection with tubercle Bacilli tends to decrease survival times.

1. Perform a two-sample t -test assuming two independent random samples from normal distributions with equal variances in SAS Studio. Report (i) the observed t -statistic and (ii) the p -value.

(i): t Value; = -3.14 (Bacilli group 1, Control group 2)

(ii): p -value: = $0.0021/2 = 0.00105$ (one-sided)

2. Perform a two-sample Welch t -test assuming two independent random samples from normal distributions with unequal variances (using the Satterthwaite approximation) in SAS Studio. Report (i) the observed t^* -statistic and (ii) the p -value.

(i): t Value; = -3.23 (Bacilli group 1, Control group 2)

(ii): p -value: = $0.0017/2 = 0.00085$ (one-sided)

T-test for Difference in Mean Times - Survival Time

The TTEST Procedure

Variable: Time

Treatment	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
Bacilli		58	242.5	117.9	15.4851	76.0000	598.0
Control		64	345.2	222.2	27.7767	18.0000	735.0
Diff (1-2)	Pooled		-102.7	180.4	32.6978		
Diff (1-2)	Satterthwaite		-102.7		31.8015		

Treatment	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
Bacilli		242.5	211.5	273.5	117.9	99.6997	144.4
Control		345.2	289.7	400.7	222.2	189.3	269.1
Diff (1-2)	Pooled	-102.7	-167.4	-37.9448	180.4	160.1	206.5
Diff (1-2)	Satterthwaite	-102.7	-165.8	-39.5736			

Method	Variances	DF	t Value	Pr > t
Pooled	Equal	120	-3.14	0.0021
Satterthwaite	Unequal	97.803	-3.23	0.0017

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	63	57	3.55	<.0001

Figure 1: Figure 1

Variable: Time							
Treatment	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
Bacilli		58	242.5	117.9	15.4851	76.0000	598.0
Control		64	345.2	222.2	27.7767	18.0000	735.0
Diff (1-2)	Pooled		-102.7	180.4	32.6978		
Diff (1-2)	Satterthwaite		-102.7		31.8015		

Figure 2: Figure 1

3. Perform a randomization/permutation test using 20,000 new random assignments of guinea pigs to treatment groups in SAS Studio. Report (i) the observed difference in sample means and (ii) the p -value.

(i): Observed Difference: = 102.7 difference in average survival times ($\mu_{\text{Bacilli}} - \mu_{\text{Control}}$)

(ii): p -value:= 0.0023 (two-sided) \rightarrow p -value:= $0.0023/2 = 0.00115$ (one-sided) using 20,000 samples and set seed 500

Note: If I'm being graded on this, please grade me on the one-sided test, as I believe that's the ultimate aim of this study.

GLM results for Difference in Mean Survival Times

The Multtest Procedure

Model Information	
Test for continuous variables	Mean t-test
Degrees of Freedom Method	Pooled
Tails for continuous tests	Two-tailed
Strata weights	None
P-value adjustment	Permutation
Center continuous variables	No
Number of resamples	20000
Seed	500

Contrast Coefficients			
Contrast		Treatment	
		Bacilli	Control
Trend	Centered	-0.5	0.5

Continuous Variable Tabulations				
Variable	Treatment	NumObs	Mean	Standard Deviation
Time	Bacilli	58	242.5345	117.9309
Time	Control	64	345.2188	222.2139

p-Values			
Variable	Contrast	Raw	Permutation
Time	Trend	0.0021	0.0023

Figure 3: Figure 1

4. Perform the Wilcoxon rank-sum test in SAS Studio. Report (i) the sum of the ranks w , for the Bacilli treatment group and (ii) the p -value.

GLM results for Difference in Mean Survival Times

The NPAR1WAY Procedure

Wilcoxon Scores (Rank Sums) for Variable Time Classified by Variable Treatment					
Treatment	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
Bacilli	58	3190.50	3567.0	195.053487	55.008621
Control	64	4312.50	3936.0	195.053487	67.382813
Average scores were used for ties.					

Wilcoxon Two-Sample Test					
Statistic	Z	Pr < Z	Pr > Z	t Approximation	
				Pr < Z	Pr > Z
3190.500	-1.9277	0.0269	0.0539	0.0281	0.0562
Z includes a continuity correction of 0.5.					

Kruskal-Wallis Test		
Chi-Square	DF	Pr > ChiSq
3.7258	1	0.0536

Figure 4: Figure 1

- (i): Sum of ranks for Bacilli $W = 3190.50$
(ii): p -value:= 0.0269 (one-sided)

5. Compare the results from parts (a)–(d).

(a): Assuming equal population variances different between Bacilli and Control groups, provides evidence to reject the null hypothesis that there is no difference in average survival times between groups of guinea pigs.

(b): Assuming population variances different between Bacilli and Control groups, provides evidence to reject the null hypothesis that there is no difference in average survival times between groups of guinea pigs.

(c): Evidence to reject the null hypothesis that the observed difference in average survival times is observed given no difference in mean survival times between groups (observed being lower average survival times for Bacilli guinea pigs compared to average Control survival times).

(d): Evidence to reject the null hypothesis that the median survival times of the Bacilli and Control guinea pig groups are the same.

Taking points (a) through (d), we find consistent evidence to support rejecting the null hypothesis of equal average and median survival times of guinea pigs between the Bacilli and Control groups. These four points provide further support of the finding that there is a difference in survival times (on average or taken as the median effect) for guinea pigs treated with Bacilli compared to control group (no Bacilli). Since the above were done using the one-sided p-value, we are rejecting the null hypothesis in favor of supporting an alternative hypothesis that the average/median survival times of the Bacilli guinea pigs is smaller/less than the average/median survival times of the Control group guinea pigs.

2.

Revisit the guinea pig study from the previous exercise. Use SAS to create diagnostic information to assess the assumptions for the traditional t -based inference procedure.

1. There is not enough information provided to assess independence of observations within each group. Which study design aspects would help you assess the assumption of independence between the groups.

There are a number of aspects of the study design that would impact whether we are violating the independence of observations in the study (both within and between groups of guinea pigs). For example, if guinea pigs were grouped together when determining which group they'd belong to, i.e., they exhibited clustered effects, then we would suspect independence being violated. Similarly, if there was some genetic relationship between guinea pigs, if they were related to one another "closely", especially if genetic relatives exhibit some similar underlying genetic condition, then independence would likely be violated. Additionally, if the same guinea pigs were measured multiple times over the course of the study, then we should suspect our assumption of independence is violated (we want to ensure that each observation is independent of every other observation, and as such want no duplicate guinea pigs in the dataset). Contrapositively, if we can ensure that none of these events/conditions **did not occur** then we would be more certain of not having violated the assumption of independence.

2. Assess the assumption of equal variances using:

1. the ratio of standard deviations;

```
knitr::include_graphics("gross.png")
```

```
191          %let a=%SYSEVALF(269.1/144.4);  
192          %put &a;  
1.86357340720221
```

```
knitr::include_graphics("RatioAssessment.png")
```

Assessing Equal Variances: **Summary Statistics**

- Check the ratio of sample standard deviations

$$\frac{\max\{S_1, S_2\}}{\min\{S_1, S_2\}}$$

- ▶ Between 1 and 2 - little impact
- ▶ Between 2 and 3 - potential impact
- ▶ Greater than 3 - likely impact

The ratio of standard deviations is given as $\frac{\max\{S_1, S_2\}}{\min\{S_1, S_{\oplus}\}}$, which from our data (shown in the above SAS output) gives us:

$$\text{Ratio} = \frac{269.1}{144.4} = 1.864.$$

Given the table from lecture for guidance, we are bordering the line of “potential impact”, but are still well within the boundary of believing this to be of “little impact” in violation of the assumption of equal variances.

2. side-by-side boxplots;

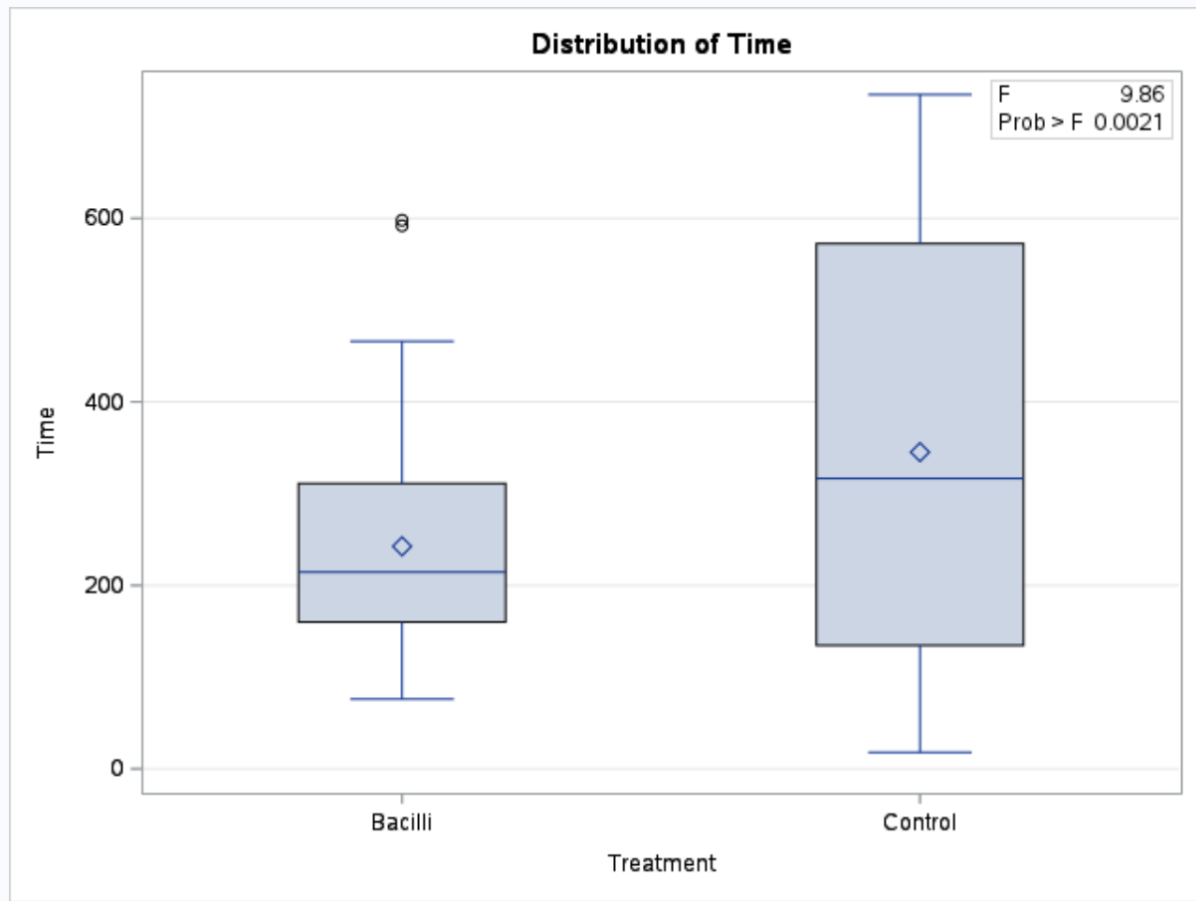


Figure 5: Figure 1

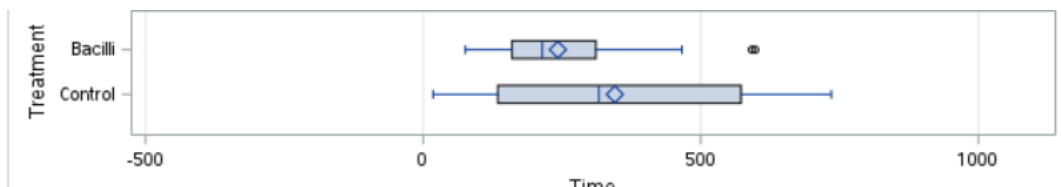


Figure 6: Figure 1

We also see a similar skewness between the two groups, insomuch as the range of values between the median and upper quartile is much wider (greater variability on the upper end of the data distribution) compared to the range of values from the lower quartile to the median. We also see the opposite of heavy tails in our data, leading to the possibility of negative kurtosis.

Overall, it appears we may have some concerns of the equal variances assumptions being valid given the visual evidence of both groups having non-normal tendencies and differences in variability (the Control group has a significantly wider range of values than the Bacilli group, Bacilli values range from 76 to 598 whereas Control group values range from 18 to 735).

3. the F-test;

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	63	57	3.55	<.0001

Figure 7: Figure 1

F-test: $= < 0.0001$, provides evidence to support rejecting the null hypothesis of equal variances, alternatively may be described as evidence to support the alternative hypothesis of a non-zero difference between the population variances of the Bacilli and Control group.

4. the Brown-Forsythe test.

GLM results for Difference in Mean Survival Times					
The GLM Procedure					
Brown and Forsythe's Test for Homogeneity of Time Variance ANOVA of Absolute Deviations from Group Medians					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Treatment	1	359581	359581	42.80	<.0001
Error	120	1008165	8401.4		

Figure 8: Figure 1

p-value: $= < 0.0001$, similar to the interpretation of the F-test p-value given above, this provides evidence to support rejecting the null hypothesis of equal variances, alternatively may be described as evidence to support the alternative hypothesis of a non-zero difference between the population variances of the Bacilli and Control group.

3. Assess the assumption of normality using the following methods:

1. side-by-side histograms;

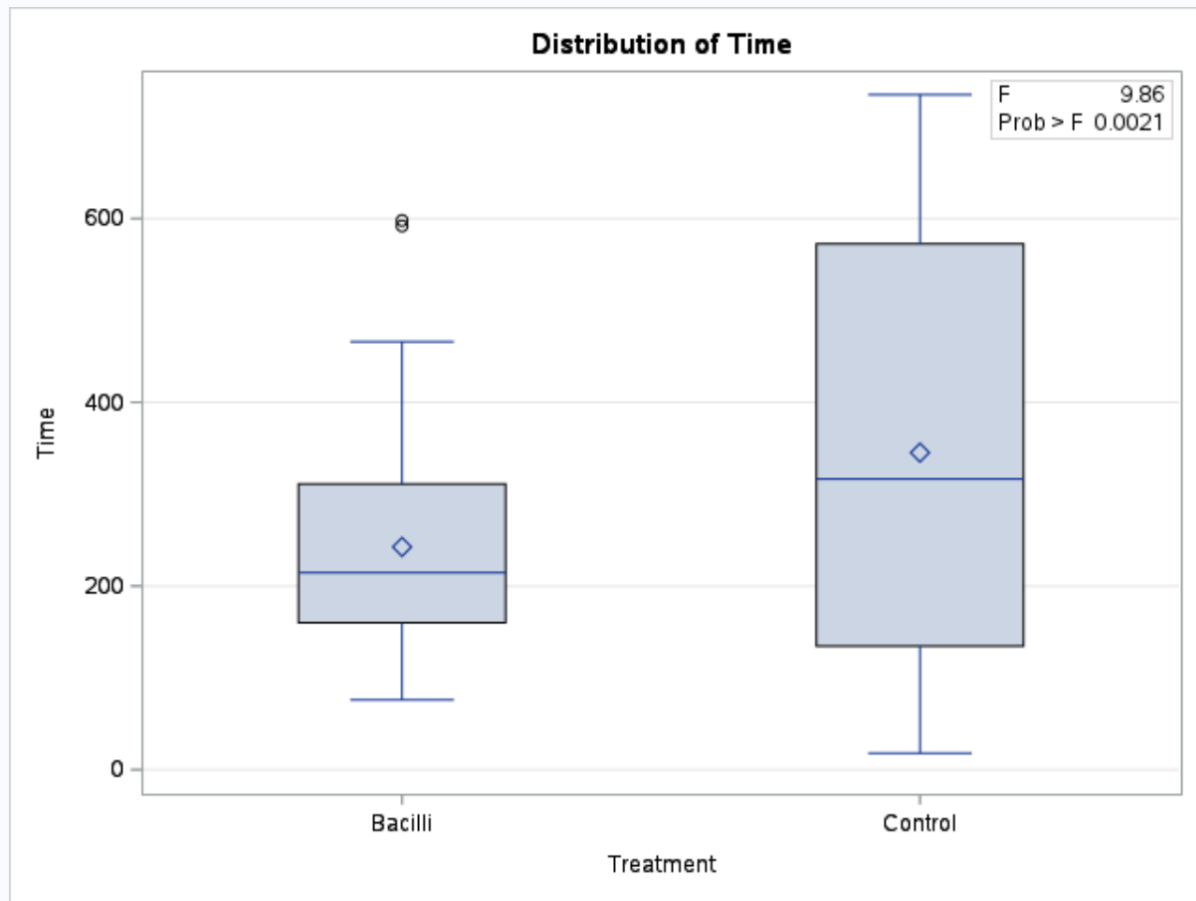


Figure 9: Figure 1

We see that for both groups the mean value is slightly greater than the median value, indicating to us that the data may not be normally distributed. Interestingly, we also see a similar skewness between the two groups, insomuch as the range of values between the median and upper quartile is much wider (greater variability on the upper end of the data distribution) compared to the range of values from the lower quartile to the median. For normally distributed data, we would expect variance to be symmetrical about the mean, which we do not observe here.

Overall, it appears we may have some concerns of the normal distribution assumptions being valid given the visual evidence of both groups having non-normal tendencies.

2. Q-Q plot(s);

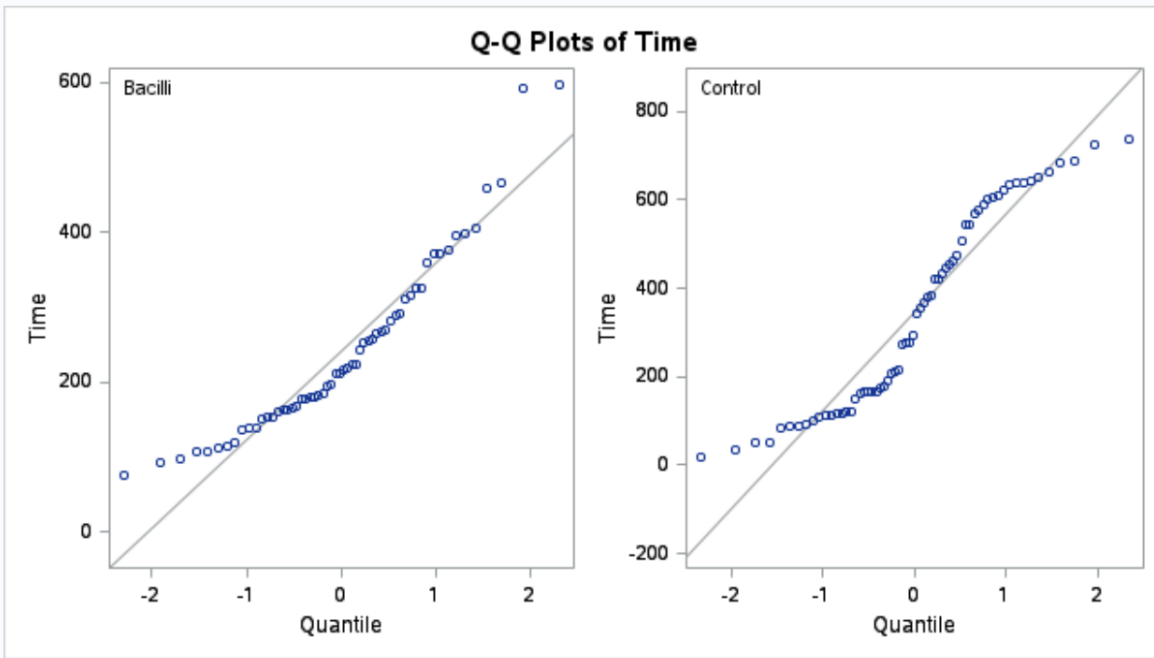


Figure 10: Figure 1

While both the distributions of Bacilli and Control groups appear similar to one another, they (both) do not appear normally distributed. Of particular note is that values of both appear to deviate around the 0-th empirical quantile, which is somewhat concerning as it is the area of the Quantile plot we'd most expect to match up to the plotted normal line. This provides evidence to possibly support the conclusion that we've violated the normally distributed observations assumption.

3. the Shapiro-Wilk test.

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.91559	Pr < W	0.0006
Kolmogorov-Smirnov	D	0.128064	Pr > D	0.0188
Cramer-von Mises	W-Sq	0.212674	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	1.317813	Pr > A-Sq	<0.0050

Figure 11: Figure 1

The above tests for normalcy provide evidence to reject the null hypothesis that our data is normally distributed. Certainly not ideal, as this is further support of the conclusion that our observations/experimental data is not normally distributed. Major bummer.

4. Based on what you learned in parts (a)–(c), complete the following exercises.

1. Summarize the assessment of all three assumptions.

Independence: We have no reason to presume one way or the other about assessing this assumption, as the assessment of this would primarily be done through evaluating the experimental design. Nonetheless, we don't have anything immediately apparent that would be a cause for concern.

Equal Variances: From both visual assessment and statistical tests, we have reason to suspect the equal variances assumption is violated and we have differences in population variance between the Bacilli and Control groups of guinea pigs.

Normally Distributed Observations (between and within groups): From both visual assessment and use of statistical tests, we have reason to believe this assumption is violated and have evidence to support the belief that our data is not normally distributed.

2. Discuss which tests are the most appropriate among the 4 used in the previous exercise.

For equal variances assumption, the F-test and Brown-Forsythe test are especially helpful and appropriate. For assessing normality assumptions, the use of the Shapiro-Wilk test is appropriate, in part given the size of our samples (Dr. Ommen made a comment that Shapiro-Wilk actually isn't great for large N as it is "bound to find something wrong with the data"). Among the visual methods, I would also argue that the quantile plot is helpful for assessing normality assumptions, though not as helpful as the Shapiro-Wilk test.

Generally speaking, the visual methods are more subjective, and lack rigor or specificity. They lead one to make assessments of the form, "it appears" or "looks like we observe", rather than evaluating some numerical value; though certainly, this may lead us to be a bit overconfident if we rely solely upon one particular numerical test.

3. Interpret the result, in the context of the study, for one of the tests you chose.

Satterthwaite Approximation: Though we have reason to suspect our assumption of equal population variances is violated, by using the Satterthwaite Approximation we are able to assess the potential difference in average survival time between the Bacilli and Control groups. When utilizing this test, we have evidence to reject the null hypothesis of average survival times between the two groups being equal. Building upon the interpretations of both the work above and prior homework, we have further evidence to support there being some effect in treating guinea pigs with the Bacilli bacteria, particularly that this treatment impacts the average survival time of guinea pigs compared to the control group (guinea pigs not treated with Bacilli).

3.

Researchers have data consisting of the annual adjusted gross incomes (**income**) for 100 randomly sampled individuals from two adjacent zip codes (**zip**). What type of statistical analysis should you use to compare the incomes in the two zip codes?

(Unsure if this is a question):

I would first test whether we can reasonably believe our base assumptions hold, and depending on assessing the validity of these assumptions proceed accordingly. If we have normality violated, then I'd recommend using a non-parametric inference technique such as Wilcoxon Rank-Sum Test. If we have normality but don't have equal variances, I'd use the Satterthwaite Approximation. And if everything assumption-wise seems appropriate I would proceed with the pooled standard deviation method with the two-sample t-test.

1. The output below was created to diagnose the assumption of normality for this data.

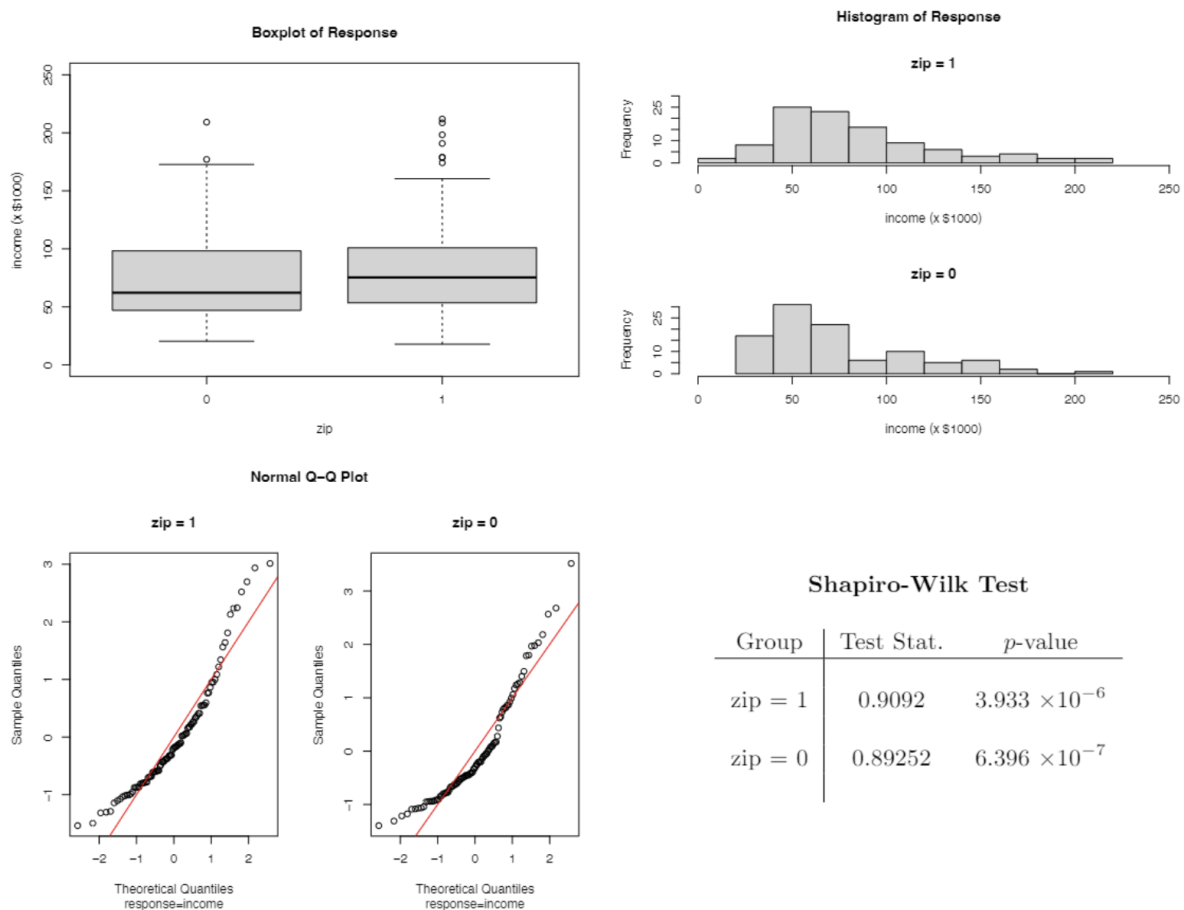


Figure 12: Figure 1

Describe three aspects (minimum) that indicate the normality assumption is violated.

(1): Histogram of Responses

The histograms above visually describe two right-skewed datasets (concentration of observations on the left-side of the distributions). By contrast, normally distributed data is not skewed, and as such we have

some semblance of (visual) evidence to suppose the underlying data used for the histograms is not normally distributed.

(2): Quantile (QQ) Plot

The Quantile Plot above shows an “S”-shaped curve in the distributions of both groups ($zip = 1$ and $zip = 0$); important for our purposes: we don’t observe many values to correspond to the “expected normal” distribution shown by the red line. This is evidence that normality is being violated in the distribution of the observations for each group.

(3): Shapiro-Wilk Test

The Shapiro-Wilk Test was done on each of the groups in the study. Both resulted in low p-values and have similarity of interpretation. Namely: These tests provide evidence to support rejecting the null hypothesis that the data are normally distributed.

2. Using the information you gathered in part (a), which transformation should be explored to achieve normality? Choose one: logarithm, square-root, arcsine, or power/Box-Cox.

Of note: We observe the values of both groups are positive. Both groups take a similar range of values, and both appear to have a right-skew. The response variable are incomes. Taking all this information into account, I would recommend using the square-root to transform the data, as it will “stretch” small values and provide greater variance. Argument against the logarithm is that we have a similar range of values between groups, and the logarithm would be more appropriate if the two groups had significantly orders of magnitude differences in incomes.

3. One of these transformations has been applied and the resulting diagnostic output is provided below.

Did the transformation succeed in remediating the non-normality?

The transformation appears to succeed in normalizing the data and remediating the non-normality shown in the first part of this problem. We see the **opposite** of what we noted in the first part: The quantile plots show observations closely plotted alongside the “expected normal” red line, the Shapiro-Wilk test has much higher p-values for each group (less evidence to support rejecting the null hypothesis of normality of the underlying distributions), as well as histograms that no longer have a right-skew (and generally appear more normal).

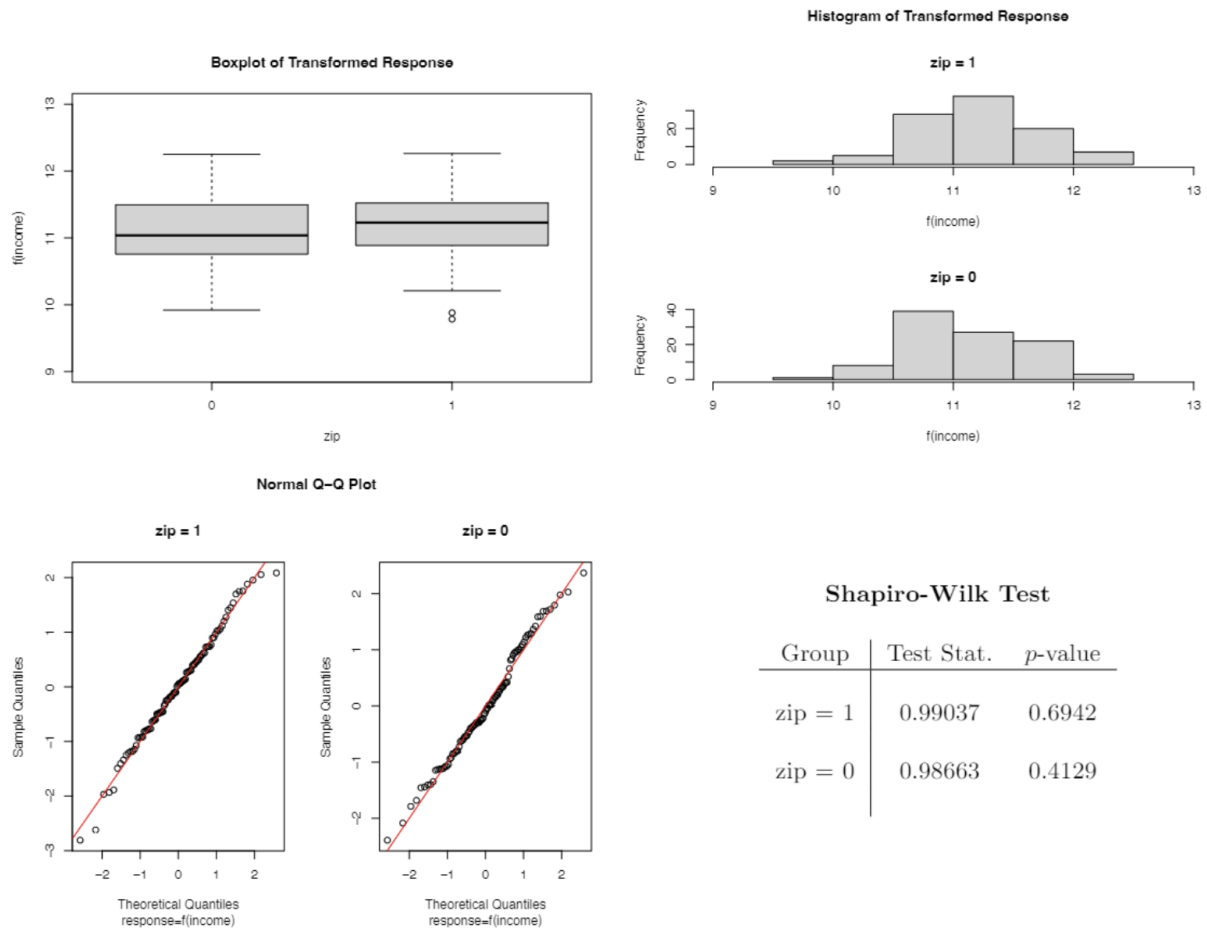


Figure 13: Figure 2

4.

Given the additional output (shown below) to diagnose the equal variance assumption, how should you proceed with the data analysis?

- Summary statistics: $s_0 = 0.4899$, $s_1 = 0.5067$, ratio = 1.0342

This provides evidence that we have similar population variances between the two groups.

- F -test: statistic = 1.0696, p -value = 0.3692

This further supports the belief of equal variances between populations, as this is a (relatively) large p -value and would provide support to not reject the null hypothesis that the two groups have equal variance.

- Brown-Forsythe test: statistic = 0.0125, p -value = 0.9109

Similar to the interpretation of the F -test, this p -value provides support in favor of the null hypothesis, not to reject it, the null hypothesis being that the two groups have the same variance.

Overall, the above would lend support to the equal variance assumption not being violated. If the variances of the populations are equal, then we can use the “regular”, i.e. not adjusted for unequal variances, two-sample t -test to access whether their average differences are equal.

Total: 50 points **# correct:** %: