

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.
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.

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.

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

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

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.
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
```

2. side-by-side boxplots;
 3. the F-test;
 4. the Brown-Forsythe test.
3. Assess the assumption of normality using the following methods:
 1. side-by-side histograms;
 2. Q-Q plot(s);
 3. the Shapiro-Wilk test.
4. Based on what you learned in parts (a)–(c), complete the following exercises.
 1. Summarize the assessment of all three assumptions.
 2. Discuss which tests are the most appropriate among the 4 used in the previous exercise.
 3. Interpret the result, in the context of the study, for one of the tests you chose.

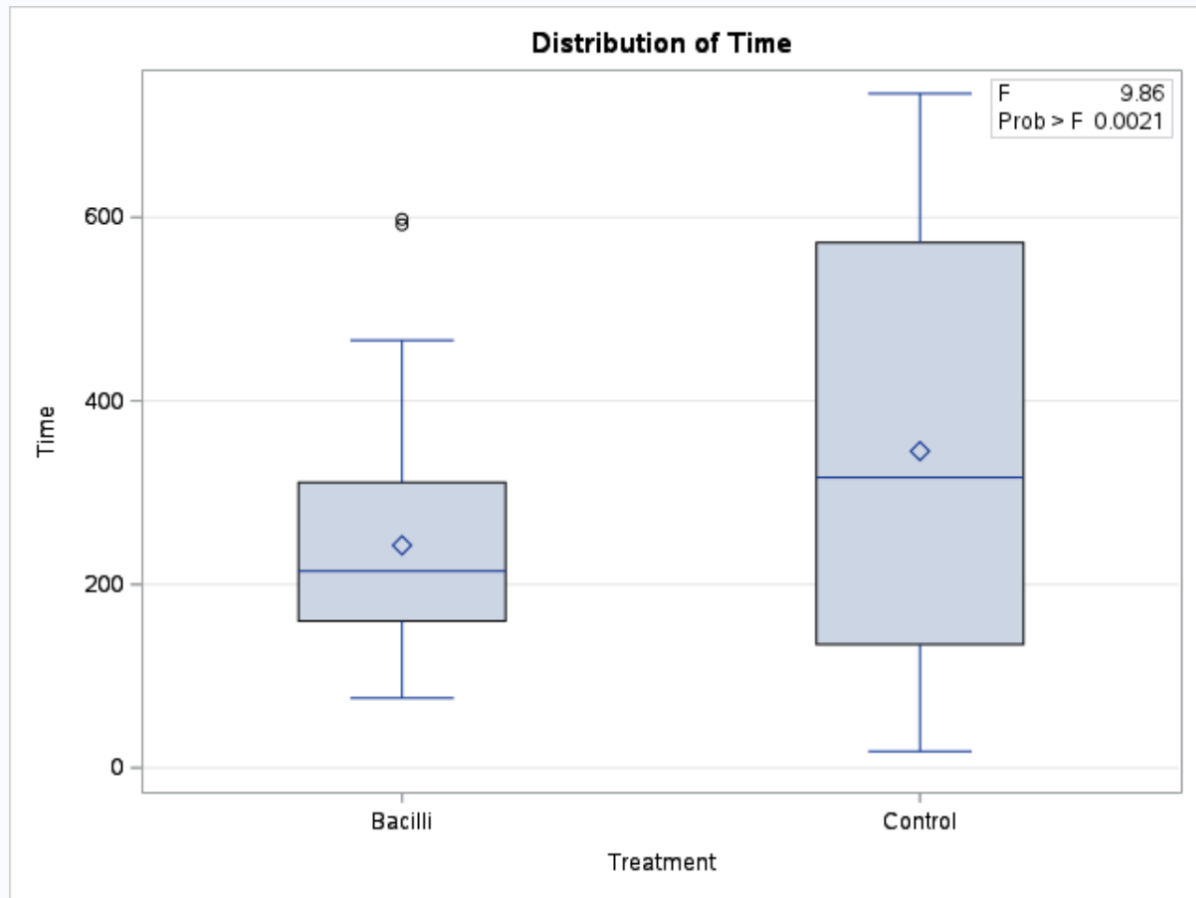


Figure 5: Figure 1

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

Figure 6: Figure 1

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 7: Figure 1

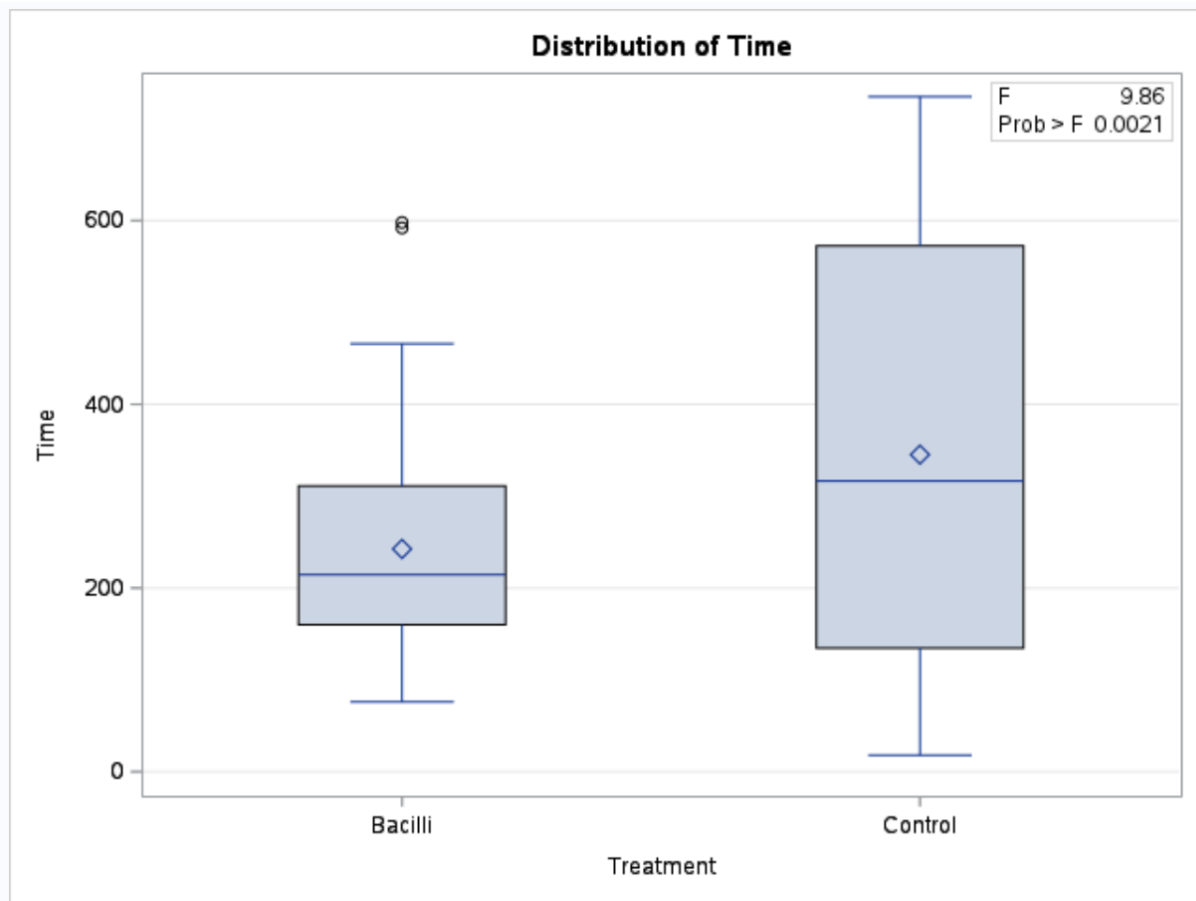


Figure 8: Figure 1

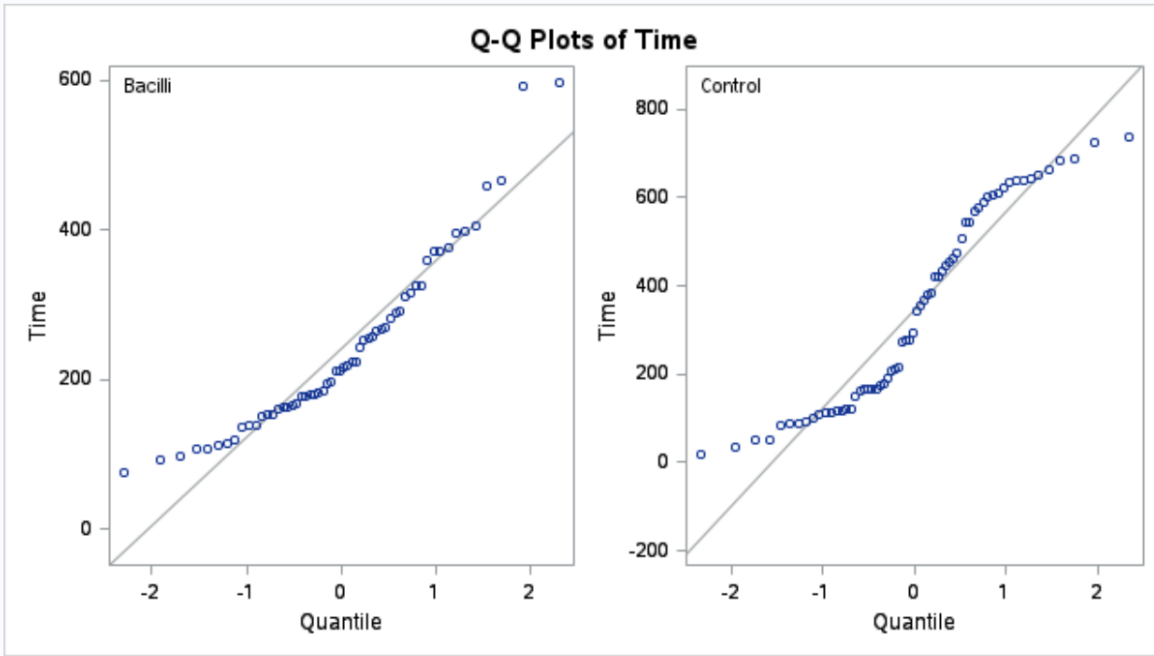


Figure 9: Figure 1

| 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 10: Figure 1

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?

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

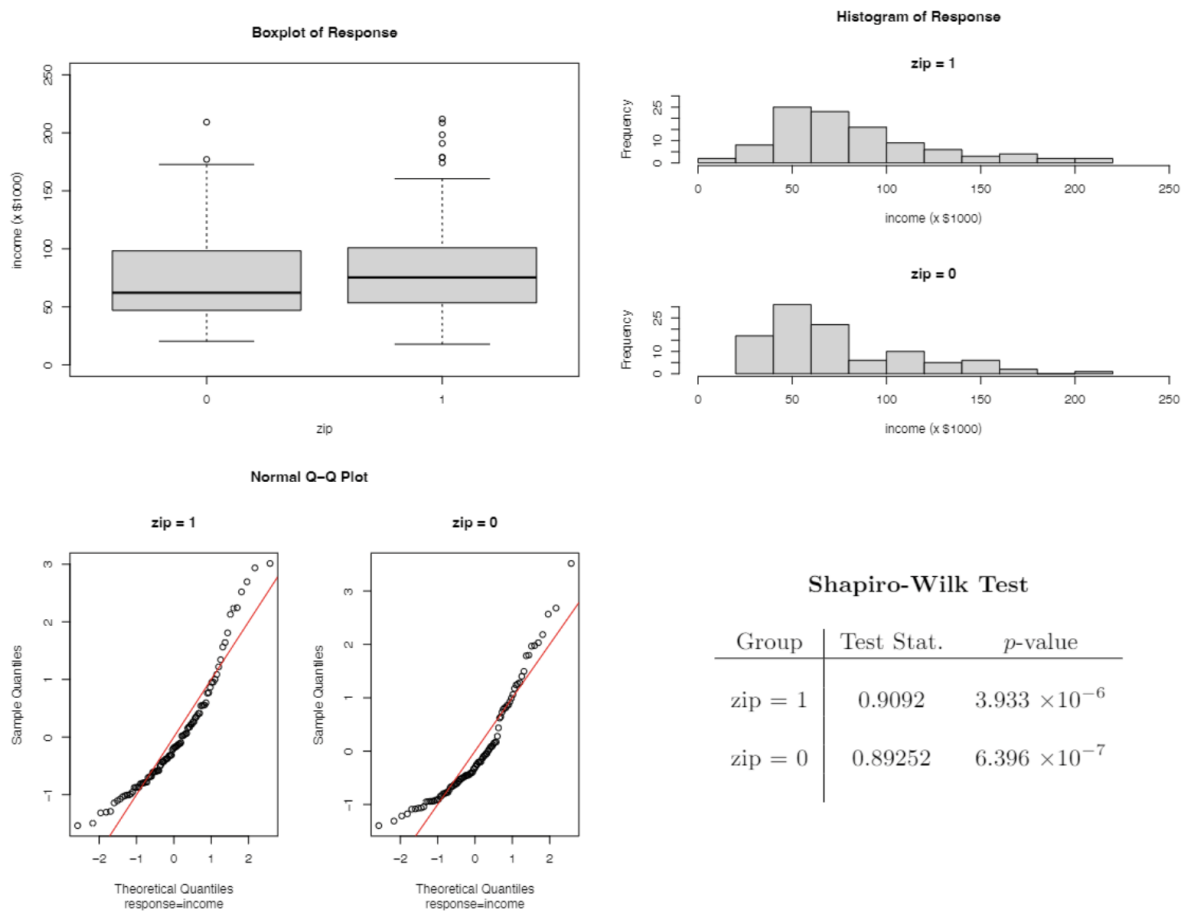


Figure 11: Figure 1

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

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.
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?

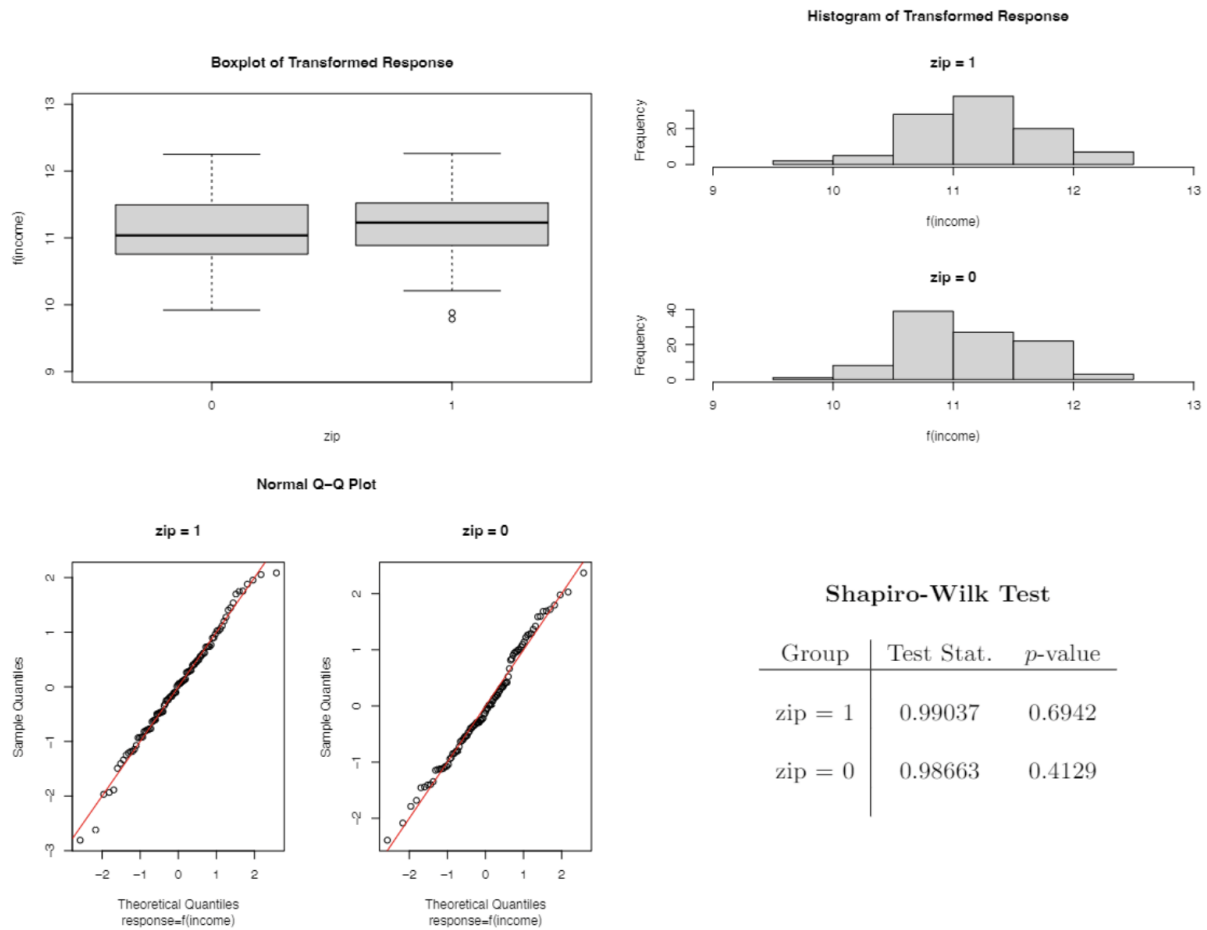


Figure 12: Figure 2

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
- F -test: statistic = 1.0696, p -value = 0.3692
- Brown-Forsythe test: statistic = 0.0125, p -value = 0.9109

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