Mini-Project 2

Q1: Groceries

1. Descriptive Summary

This data set represents a sample of 29 low-price grocery staples from a WinCo and Walmart store in January of 2016. The main question at-hand is "Do the data suggest that prices tend to be higher at one of the chains?" This data set contains only three columns: the name of the item (*Item*), the price of this item at WinCo (*WinCo*), and the price of this item at Walmart (*Wal.Mart*). Because the data in the *Item* column is of the character data type, a univariate summary could not be provided. However, I was able to calculate the univariate summaries for the latter two fields. *WinCo* had a minimum value of 0.420, a first quartile of 1.01, a median of 1.68, a mean of 2.479, a third quartile of 2.77, and a maximum of 12.68. *Wal.Mart* had a minimum of 0.56, a first quartile of 1.23, a median of 1.77, a mean of 2.659, a third quartile of 2.74, and a maximum of 12.84. For comparative visuals of the univariate summaries, please refer to Figures 1, 2, and 3. Looking at Figure 1, it is evident that there exists a strong, positive relationship between the pricing at these two grocery stores. The Pearson correlation coefficient between the two is 0.9904, which is not only indicative of a strong, positive association, but also of a nearly-perfect association. Looking at the boxplots in Figure 2, there does not appear to be much variation between the two; at first glance, they actually look identical! The final descriptive summary I performed was taking the differences between the two fields. This was done by subtracting the price of a given item at Walmart from the price of that same item at WinCo. The summary of the differences yielded a minimum of -0.95, a first quartile of -0.45, a median of -0.09, a mean of -0.1799, and third quartile of 0, and a maximum of 0.5 (Figure 3).

Figure 1: Scatterplot of Prices

Figure 2: Boxplot of Univariate Summaries

Figure 3: Boxplot of Differences in Prices

2. Inferential Analysis

For the inferential analysis, I decided to conduct both a confidence interval and a hypothesis test. Before conducting these two tests, I first assessed whether the data was paired or independent. After some analysis, I concluded that the data was paired, based upon the way in which the WinCo and Walmart prices corresponded to one another, side-by-side, in the data frame. When creating the confidence interval, I decided to use a confidence level of 95%. The results of the t-test from Figure 4 provided a confidence interval of (-0.31527283, -0.04458924), which means there is a 95% chance that the mean difference in pricing between WinCo and Walmart is between -0.31527283 and -0.04458924.

When conducting the hypothesis test, I also decided to use a confidence level of 95%. The null hypothesis was $H_0 : \mu 1 - \mu 2 = 0$, the alternative hypothesis was $H_A : \mu 1 - \mu 2 \neq 0$, the test statistic was -2.7233, and the p-value was 0.011. Based on the results of the test, we reject the null hypothesis because the p-value of 0.011 is less than the alpha of 0.05. There is sufficient evidence to suggest that the means between the two grocery stores are indeed different.

```
> #Confidence Interval and Hypothesis Test
> t.test(grocery$differences, conf.level=0.95)

        One Sample t-test

data:  grocery$differences
t = -2.7233, df = 28, p-value = 0.011
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -0.31527283 -0.04458924
sample estimates:
mean of x
-0.179931
```

Figure 4: Confidence Interval and Hypothesis Test for Differences in Pricing Between WinCo and Walmart

3. Conclusion

The data does indeed suggest that prices tend to be higher at one of the chains. The 95% confidence interval of -0.31527283 and -0.04458924 supports the idea that the two means are not equal, as the entire interval remains below zero. The hypothesis test's ruling to reject the null hypothesis also supports this conclusion. Looking at the normal quantile plot in Figure 5, the pattern of the plot is very close to the line, so it is reasonable to assume normality in the data.

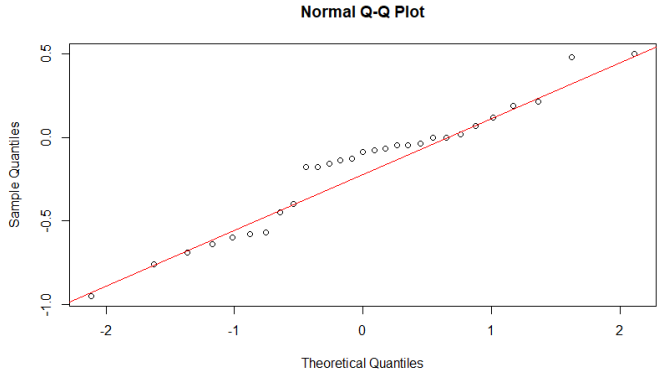

Figure 5: Normality of the Data

There are a few outliers that exist in the data. Looking at the boxplots in Figure 2, one can see six outliers total (three from each store). These outliers cause the average price to be positively skewed. This positive skew can also be identified when comparing the means of 2.479 and 2.659 to the medians of 1.86 and 1.77, for WinCo and Walmart, respectively. I do not believe that these outliers had a material impact on the inferential analysis, especially considering the pricing data for WinCo and Walmart was so highly correlated. Even if there were not outliers, the outcomes of the confidence intervals and hypothesis tests would have likely remained the same.

Q3: Teen Dating Violence

1. Descriptive Summary

This data represents a survey of 580 high school students in South Carolina, which asks whether they experienced any dating violence from grades 9-12. For this analysis, the main question to consider was "Were students of a particular ethnic group more likely to be lost to follow-up?" To keep things simple, I decided to compare survey follow-up between two of the three reported ethnicities: blacks and whites. Figure 6 provides a matrix of the response rates across these two ethnicities. Among the surveyed black seniors, 48 were lost to follow-up, compared to just 39 for white seniors. Figure 7 converts these frequencies into percentages. According to the proportion table, approximately 16.78% of the surveyed black high school seniors did not follow-up, compared to 17.03% for white high school seniors.

	Lost to Follow-Up	Follow-Up
Black	48	238
White	39	190

Figure 6: Matrix of Lost to Follow-Up and Follow-Up by Ethnicity

	Lost to Follow-Up	Not Lost to Follow-Up
Black	0.1678322	0.8321678
White	0.1703057	0.8296943

Figure 7: Proportion Table of Lost to Follow-Up and Follow-Up by Ethnicity

2. Inferential Analysis

For this inferential analysis, I decided to conduct both a confidence interval and hypothesis test, using a 95% confidence level for each. The results for the confidence interval were (-0.07011030, 0.06516329). This means that there is a 95% chance that the mean difference in survey follow-up between black and white high school seniors is between -0.07011030 and 0.06516329.

Regarding the hypothesis test, the null hypothesis was $H_0 : \pi1 - \pi2 = 0$, the alternative hypothesis was $H_A : \pi1 - \pi2 \neq 0$, the test statistic was 1.626635e-15, and the p-value was apparently 1. Based on the results of the test, we fail to reject the null hypothesis because the p-value of 1 is greater than the alpha of 0.05. Sufficient evidence does not exist to suggest that the mean difference in survey follow-up between black and white high school seniors is different.

```
        2-sample test for equality of proportions with continuity correction

data:  survey_table
X-squared = 2.6459e-30, df = 1, p-value = 1
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.07011030  0.06516329
sample estimates:
   prop 1    prop 2
0.1678322 0.1703057
```

Figure 8: Confidence Interval and Hypothesis Test for Difference in Follow-Up Between Black and White Students

3. Conclusion

The data does not suggest that students of a particular ethnic group are more likely to be lost to follow-up. This conclusion is supported by the 95% confidence interval of (-0.07011030, 0.06516329), as well as by the hypothesis test's conclusion to fail to reject the null hypothesis. I believe this data comes from a well-defined population. Based upon the large sample size of 580 high school students, it is fair to assume normality in the data. Other than the nonresponse bias associated with "lost to follow-up," there does not appear to be any major irregularities with the data. Interestingly enough, this nonresponse bias was actually necessary for the study, as the major question at-hand called for the comparison of lost to follow-up rates between two of the surveyed ethnic groups.