Project 1 – Part 1

Ian Sims

February 6, 2019

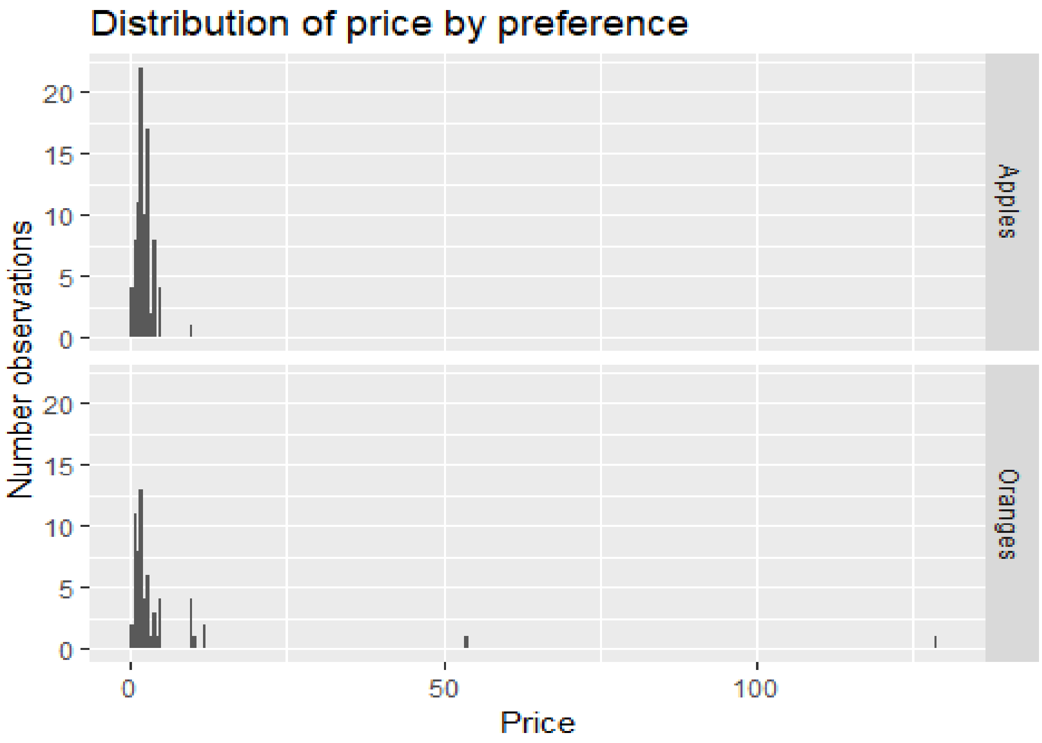
Determining the Sample Size

The first step is to estimate the sample size that will be required to meet the desired Power (80%). We can estimate this sample size by utilizing the existing sample. In this way we will consider the first sample our pilot data. A note about the pilot sample data

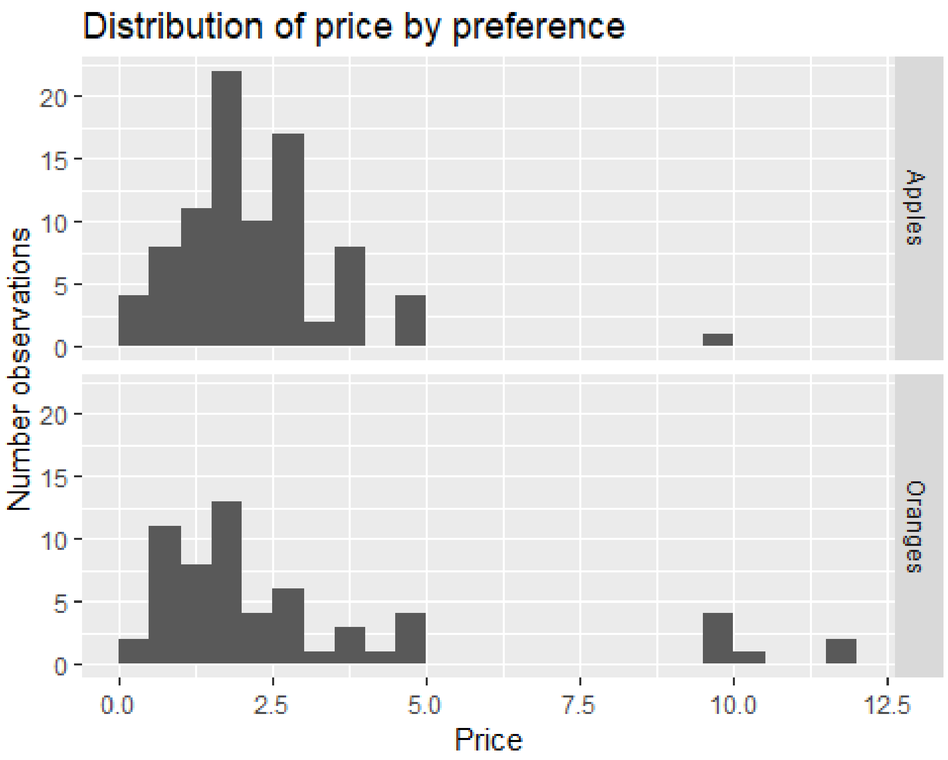
A note about the pilot sample data

When reviewing the sample data provided in the oranges.txt file, several significant outliers are observed. For oranges there are 63 observations. Most of the prices listed are under 10.00, with multiple observations at or near 10.00. However, there is one observation for 1000.00, one for 129.00, and one for 54.00. If these values are included, the mean price for the oranges group is: 21.79. If these three observations are removed the mean price for the orange group falls to approximately: 3.16. These outliers have similar large impacts on the calculated standard deviation.

The impact of these outliers can be seen in the histogram below for the 'Oranges' preference.



Removing the outlier data points by restricting price data to be less than 50.00, produces a much more reasonable distribution as can be seen below.

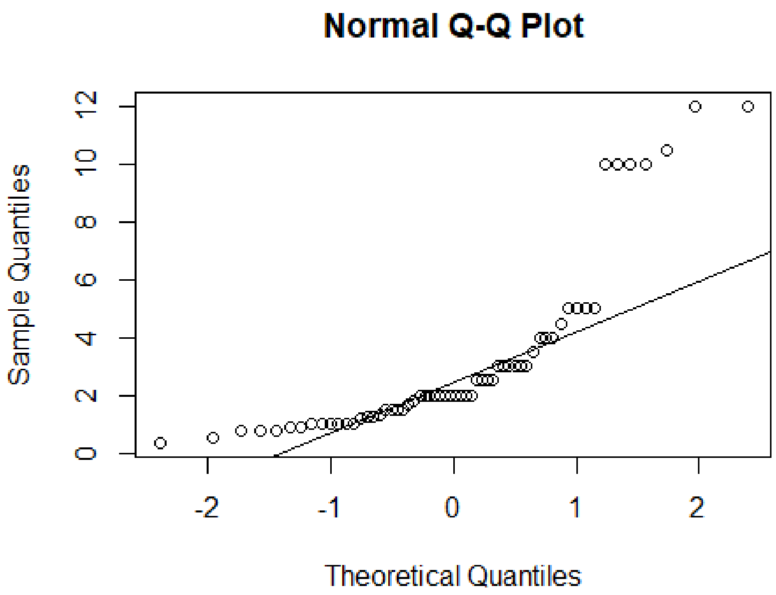


Based on the significance of these three extreme outliers, for the purposes of calculating an appropriate sample size we have removed them from the ‘Oranges’ pilot data.

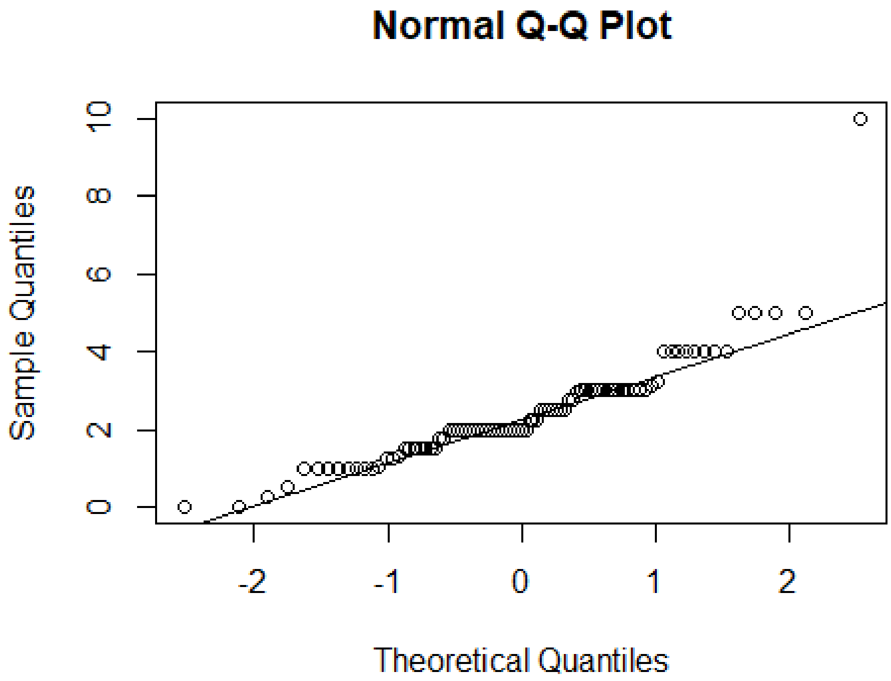
Normality assumptions

Many statistical tests applied to hypothesis tests make assumptions that the underlying distributions being studied are normally distributed. This would include the T-test. As can be seen from the above histograms, this assumption does not seem to hold for the pilot data. Viewing the QQ-plots for both apples and oranges (below) further suggests that the normality assumption doesn’t hold.

QQ Plot for oranges



QQ plot for apples



This lack of normality leads us to consider nonparametric tests. The benefit of a nonparametric test is that it is based on fewer assumptions, including the assumption of normality. For the purposes of calculating an appropriate sample size we used the pilot data in conjunction with a nonparametric test known as the Wilcoxon test. We used this test in simulation assuming an underlying exponential distribution based on the means in the pilot data. We ran this simulation testing for a Power of 80% and a 5% level of significance. Based on this simulation the needed sample size to produce an 80% power for each of the preferences is approximately 225.

If we assume that the proportion of preferences in the pilot data is representative of the general population, then the total of observations we would need to collect to arrive at at least 225 for each preference is approximately: 600.

Recommended Plan

#1 Collect a sample with size approximatley equal to 600.

Care should be given when collecting this sample to ensure that it is drawn from a random population.

#2 Evaluate the data for outliers and to determine if normality assumptions are reasonable.

If the new sample does not appear to be normally distributed as the pilot data suggests then nonparametric tests should be considered.

#3 Perform hypothesis test.

As mentioned, if the distribution doesn't appear normal (like the pilot data) perform a hypothesis test using a nonparametric test. The null hypothesis for this test would be that the mean price between the different preferences are the same. The alternate hypothesis would be that the mean price for oranges is larger than for apples.

One nonparametric test that could be used is the Wilcoxon test. This test would rank the prices for each preference and would use the differences in the sum of the ranks for the test statistic. This would be a good test because it is not as sensitive to skewed data (as the pilot data appears).