Project 3: Population Proportions

Zack Garza

Math 142: Elementary Statistics
Effective Date of Report: July 19, 2014

Contents

I. Introduction

THe objective of this project is to test a claim about two independent population proportions. Due to the ever-increasing cost of fuel and the intricate link between fossil fuels and international conflicts, the overall fuel efficiency in miles-per-gallon (MPG) has become increasingly important to both consumers and manufacturers. The question addressed in this project is whether or not spikes in global oil prices drive innovation by increasing the average miles per gallon rating in newly manufactured cars.

II. Data

To analyze this hypothesis, we turn to the UCI Machine Learning Repository, which provides a data set of matching vehicles made between 1970 and 1982 and their efficiencies in MPG. The time period between 1978 and 1982 will be examined, as this time interval includes the 1979 Energy Crisis, during which the US price per barrel of crude oil spiked and purportedly led to increased fuel economy in the 1980s.

The data set includes samples of individual car models from each year, so the first population will be a sample of cars automobiles made in 1978, and the second population will be those made in 1982. The 10 year average MPG between 1970 and 1980 will be used as a baseline – anything above that average will be considered a "success", and anything below it will be considered a "failure". We will take the null hypothesis to be that the proportion of 1978 cars that were above the ten year average is equal to the the same proportion of 1982 cars. The alternative hypothesis is that they were not in fact equal, which is the primary claim we seek to, and is summarized as:

$$H_0 : p_1 = p_2$$
$$H_1 : p_1 \neq p_2 \text{(Primary Claim)}$$

where p_1 is the proportion of 1978 cars for which the MPG is above the 10 year average, and p_2 is the proportion of 1982 cars with MPGs above the 10 year average.

First, the baseline must be computed, so the data set is read in, and the MPGs for all cars made between 1970 and 1980 are summed and divided by the number of cars sampled.

Ten Year Average: 23.515

Before performing the analysis, we check that the requirements for testing two sample proportions are met.

1) We first assume that the data collected constitute simple random samples.
2) Since the two samples are disjoint and not related in any immediate way, we assume that the samples are independent.
3) Both samples also exhibit 5 success and 5 failures, such that $n\hat{p} \geq 5$ and $n\hat{q} \geq 5$.

Since the requirements are met, statistical characteristics of these samples are calculated and listed below.

Population	Number of Samples (n)	Average MPG (μ)	Proportion Above 10-Year Average ($\hat{p}$)	Standard Deviation (s)
1 (1978)	$n_1 = 36$	$\mu_1 = 24.061$	$\hat{p}_1 = 0.389$	$s_1 = 6.898$
2 (1982)	$n_2 = 31$	$\mu_2 = 31.71$	$\hat{p}_2 = 0.935$	$s_2 = 5.393$

The pooled sample proportion $\overline{p}$ is 0.642, which is given by the equation

$$\overline{p} = \frac{x_1 + x_2}{n_1 + n_2}. \tag{1}$$

The test statistic z is -4.648, which is given by the equation

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{\overline{p}\overline{q}}{n_1} + \frac{\overline{p}\overline{q}}{n_2}}}, \tag{2}$$

where $p_1 - p_2$ is assumed to be zero, according to the null hypothesis. In other words, the null hypothesis assumes that the 1978 and 1982 populations are equal.

Because the alternative hypothesis involves inequality, the p value is obtained from a two-tailed test based on the z-score from above. We therefore look for the probability of getting a z-score at least as extreme as the test statistic. For this test, we use a significance level of $\alpha = .01$, which corresponds to the critical values $z_{\alpha/2} = \pm 2.575$. Plotting the normal distribution shows where the test statistic lies in relation to the critical values:

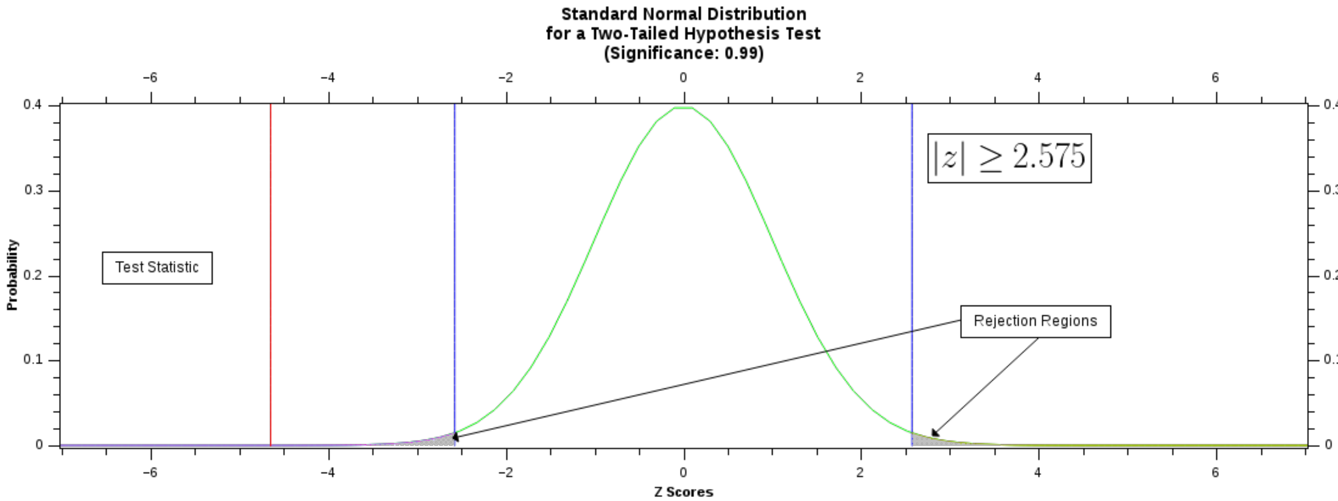

Figure 1. Representation of critical values, the location of the test statistic, and the regions in which the null hypothesis would be rejected at a significance level of 0.99.

By calculating the area under the normal distribution, we find that the cumulative area in the tail of the test statistic is 1.676e-06, and that the area in both tails is equal to 3.352e-06.

Thus, we have

$$p = 3.352\text{e-}06,$$

and since p is much lower than the significance level, we reject the null hypothesis and conclude that there is sufficient evidence to support the claim the the population proportions are equal.

Based on these results, we find that the margin of error is 0.238, and thus the confidence interval is

$$-0.784 < (p_1 - p_2) < -0.308.$$

Since the confidence interval does not include zero, we have reasonable evidence to suggest p_1 and p_2 are not equal.

III. Conclusion

Because H_0 was rejected, we can conclude that there is sufficient evidence to suggest that there was a spike in the average fuel economy between 1978 and 1982 compared to the average fuel economy of cars made in the 1970s. It can not be definitively concluded that this was caused by the spike in oil prices that occurred in the late 1970s, statistical analysis suggests that cars made in the early 1980s are from a population with fuel efficiency that was much higher than the average at the time.