the steele group
experience
knowledge
perspective℠

Using Kruskal-Wallis to Improve Customer Satisfaction

A White Paper by

Sheldon D. Goldstein, P.E.
Managing Partner, The Steele Group

Using Kruskal-Wallis to Improve Customer Satisfaction

KEYWORDS

Kruskal-Wallis Method
Equality of Means
Customer Driven Quality
Voice of the Customer
Quality Improvement

1. INTRODUCTION

There is no benefit in improving business attributes that are not likely to measurably increase customer satisfaction. The use of the “Kruskal-Wallis one-way analysis of variance by ranks” to compare the means of satisfaction scores of several company attributes that are important to your customers can focus attention on those attributes in need of the most improvement. A statistical approach can provide confidence in identifying valid problems and recommending solutions that have a better chance of resulting in improved satisfaction.

Many customer satisfaction programs start and end with a measure of customer satisfaction and a pledge to improve in the future. It is much harder, and perhaps less common, for company personnel to evaluate the details of customer feedback surveys, determine which company attributes may result in increases in customer satisfaction (if improved), and then set out to implement changes that are recommended by your customers.

However, just because a company attribute has a low customer satisfaction measure doesn’t mean it differs statistically from the other attributes important to your customers. Choosing the right ones to improve maximizes the use of resources and has a better chance of a positive return on your efforts.

1.1 Scope of the Project

The customer satisfaction issue analyzed in this article uses data that was taken from a full year of customer satisfaction surveys from a company that provides commercial plumbing products and services to businesses in a Midwestern city. Customers were surveyed to choose those attributes they thought were most important when considering service from the company. They said that the most important attributes this company could provide to meet their needs and generate high levels of satisfaction are:

- Courtesy
- Scheduling a convenient time for service
- Arriving when promised to perform the service
- Meeting customer expectations

- Being neat and clean
- Answering questions about the service when asked
- Competitive pricing
- Quality of the products
- Providing overall value

Using this information, customer satisfaction surveys were mailed each month to all customers who purchased products or services, to assure each survey corresponded to a current service. This minimizes the issues of customers having to recollect a service that occurred a long time ago, or consolidate their feelings for several services. If we are to take control over declining customer satisfaction, we want timely feedback.

The company enjoyed a high response rate of 20% on survey returns. However, that means that 80% of customers did not respond. Recognizing that initial non-respondents to the feedback survey tend to have different opinions about the service than first-time respondents, and therefore they display different customer satisfaction metrics, considerable effort was spent to solicit and receive first-time non-respondent survey information to round out the information profile. Initial non-respondents were contacted again, 30 days later, with a repeat request for feedback.

In all, there are 517 first-time respondent surveys included in this analysis and 139 first-time non-respondents who eventually sent in survey results. These two groups were separated because they are considered to have different opinions of the company's performance in each attribute, and we wanted to see if there are indeed any differences in the conclusions we would measure from these groups.

1.2 Methodology

The premise of this article is that we can take a statistical approach to analyze the details of customer satisfaction surveys and determine which attributes are most favorably scored by customers and which ones need improvement.

We must be careful every time we draw statistical conclusions from data. In this case, there are important issues because of the nature of satisfaction data in general.

1.2.1 Statistical Considerations

We have compiled measures of performance for each of the nine attributes. We asked respondents to rate the attributes on a scale of 1 to 10. This represents an ordinal scale. Ordinal scales do not have meaningful intervals. For instance, the difference between a satisfaction improvement from 5 to 6 does not necessarily require the same level of effort as an improvement from 8 to 9. Neither is there a ratio scale. A score of 8 is not twice the level of satisfaction of a score of 4.

For these reasons, it is not strictly appropriate to use standard statistical methods to evaluate customer satisfaction metrics if the underlying probability distribution isn't normal. However, this is often done. One rationale for using statistical metrics, such as mean and standard deviation for ordinal data is that we understand the limitations of interpreting the results. Another is that we usually have large sample sizes, as in this case where we have 517 respondent surveys in our sample. Finding the mean of hundreds of responses, despite the lack of normality, gives us a measure of the average value of satisfaction.

However, the non-normality of the underlying population brings discomfort to those who want to meet the requirements of a symmetric population. In satisfaction data we rarely have a normal distribution. Therefore, we look toward the Kruskal-Wallis test, which gives us a way to evaluate ordinal data in more depth and draw strict statistical interpretations from the results, including comparison of means.

1.2.2 Statistical Significance and Random Variation

We ask whether there is a difference between attribute means from a statistical perspective because we want to know if we are just seeing random variation in the means, or a variation that is the result of a special cause. For instance, let's say that the mean of a satisfaction score is 9.05. That represents the mean of 517 respondents who scored the attribute. If we were to compile another set of scores at the same time and of the same population, in all likelihood, a different sample of respondents would send in surveys. Some respondents would be the same and some would be different, but the sample would differ from the one summarized in this analysis. Since both samples would be measuring the same population at the same time, any difference in mean scores would be caused by random variation in sampling. We would not conclude that there is a statistically significant difference in those scores.

However, if we can assess whether the variation between the means of the different attributes is the result of random variation, or whether it is the result of a more serious cause (as confirmed by the statistical analysis), then we can work on those attributes that show enough difference from the overall responses to warrant action. Working on attributes whose lower scores are the result of random variation of sample data will not result in meaningful improvement in customer satisfaction.

1.2.3 Independence of the Samples

While the Kruskal-Wallis test can be performed on data that is not from a normal distribution and is ordinal in nature, it requires that the samples are independent. The definition of independence requires that the scoring of one attribute would not influence the scoring of another attribute. This does not mean that there can be no correlation between attributes, only that the answers don't influence each other. For example, if we advertise heavily for a product, and sales increase, we may conclude that there is a correlation, and further that increasing advertising influenced sales. These events would not be independent. However, the advertising may have been totally ineffective, and sales increased because prices were drastically reduced at the same time that the advertising campaign was launched. In this case price and sales are correlated and these events are not independent, but advertising and sales, while correlated, are really independent.

In this case, customer responses to the attributes “did we arrive on-time” and “quality of service”, for instance, should not influence each other. Even the attributes that are related, (“price” and “value”) share so little in common that we assumed that there is independence between those two attributes as well. There have been many cases where price was scored low and value was scored high.

2. KRUSKAL-WALLIS METHODOLOGY

The “Kruskal-Wallis one-way analysis of variance by ranks” is a method of comparing different samples to calculate whether there is a statistically significant difference between the ratings of those attributes. The method relies on the ranks of the scored values and the means of those ranks, rather than examining the means of the data.

Just because there is a difference in the averages doesn’t permit us to conclude that the difference is statistically meaningful. Once we decide how sure we want to be about our conclusions, in this case we are using a .05 significance level (95% level of confidence), we conduct the Kruskal-Wallis test to decide if any attributes are statistically different from the others with the specified degree of significance.

First, we set up a simple hypothesis test that postulates there is no difference between the satisfaction scores of any of the attributes. The null hypothesis is Ho: and the research, or alternate hypothesis is Ha:.

Then:

Ho: all attribute populations are identical
Ha: all attribute populations are not identical

We are assuming that there is no statistically significant difference between the means in the null hypothesis (Ho). When we employ the Kruskal-Wallis test statistic, we are testing the validity of this hypothesis. This test relies on a distribution that is approximated by a chi-squared distribution with degrees of freedom k-1, or the number of attributes being compared minus 1. The test ranks responses based on the raw data (scale of 1 to 10 responses). For an in-depth review of the method, you may find it in the reference text listed at the end of this article.

The ranking of responses is performed by setting-up a rank for all nT data points and then summing the ranks of the data in each sample. We then calculate the test statistic:

$$W = \left[\frac{12}{n_T (n_T + 1)} \sum_{i=1}^{k} \frac{R_i{}^2}{n_i} \right] - 3 (n_T + 1)$$

Where:

k = the number of attribute samples
ni = the number of responses in sample i
nT = the total number of responses in all samples
Ri = the sum of the ranks for sample i

2.1 Corrections for Tied Observations

In this case, there is one further factor to consider, and that is ties in the ranks. This is referenced in most books that cover tests of ranks. Whenever the data have repeat scores, for example a rating of 9 from many respondents, the ties must be considered in a correction factor applied to the value of W to account for the effect of ties.

The correction factor is:

$$C^* = 1 - \frac{\sum_{i=1}^{e} (t_i^3 - t_i)}{n_T^3 - n_T}$$

Where:

e = the number of different observations in the samples
ti = the number of observations tied with the ith observation in size
nT = the total number of responses in all samples

Then,

W corrected = W/C*

The test is the same as any comparison of means. Whenever the value of W corrected is greater than the chi-squared table value at degrees of freedom (k-1) and the specified significance level, then we reject the null hypothesis and conclude that the means differ. When W corrected is less than the table value, we conclude that the means are statistically equal, in other words, there is no reason to believe that the attributes differ in their ratings from a statistical perspective.

See Figure 1, which is a graph of the chi-squared distribution showing the .05 significance level for the hypothesis test.

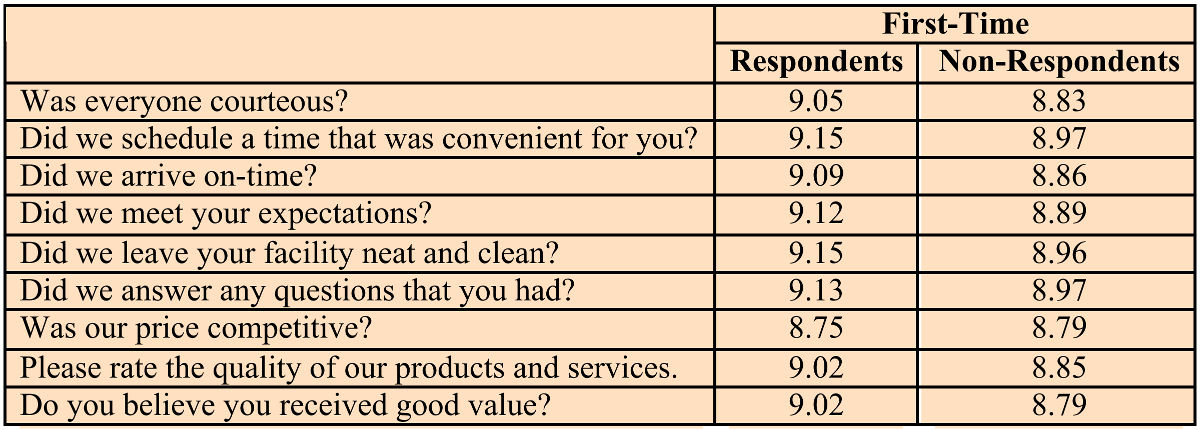
Once the conclusion is made that there is a difference in means, we conclude that the mean that stands out as the highest or lowest is the one that is statistically different.

3. RESULTS

3.1 Data Summary

When all the raw data were compiled, we calculate an “average satisfaction score” for each attribute to see if we could notice any differences in perception reported by customers for each attribute.

On a scale of 1 to 10, with 1 being “Not at all Satisfied” and 10 being “Exceeded my Expectations.” The average results are:



	First-Time	
	Respondents	**Non-Respondents**
Was everyone courteous?	9.05	8.83
Did we schedule a time that was convenient for you?	9.15	8.97
Did we arrive on-time?	9.09	8.86
Did we meet your expectations?	9.12	8.89
Did we leave your facility neat and clean?	9.15	8.96
Did we answer any questions that you had?	9.13	8.97
Was our price competitive?	8.75	8.79
Please rate the quality of our products and services.	9.02	8.85
Do you believe you received good value?	9.02	8.79

This is the natural way we would choose to analyze the data. For the first-time respondents, the obvious outlier is price, with a score of 8.75, much lower than any other attribute score. And, it would be easy to conclude that customers believe price is too high, resulting in the lowest satisfaction score amongst all the attributes. When looking at the first-time non-respondents, this conclusion is not so obvious. While scores for almost every attribute are lower in this group than for the initial respondents, price did not stand out as being much lower than the other attributes. And, it is higher than the score from first-time respondents.

The next step in the process is to conduct a formal analysis to study whether there is a difference based on the Kruskal-Wallis test.

3.2 Kruskal-Wallis Conclusions for Initial Respondents

The results show that price is an issue with customers. In fact, at the .05 significance level and for 8 degrees of freedom (9 attributes),

$$W = 23.97$$

where the test disproves equality of the means anytime the calculated value of the test statistic (W corrected) is greater than 15.51. This is conclusive evidence that at least one of the attributes is statistically higher or lower than the other attributes. In the calculation of the value W corrected, we must evaluate the sum of the ranks for each attribute, which is the measure of the composite score for each of the attributes. Price has the lowest rank measure of all the attributes, and it confirms the relative standing of price in the means of the attributes calculated earlier.

See Figure 2 for a graph of the average satisfaction scores.

3.3 Kruskal-Wallis Conclusions for Initial Non-Respondents

We wanted further validation that price is a real concern to our customers. Recognizing that the sample of customer responses represented a healthy response rate of 20% of all surveys mailed, still 80% of the customer base did not send back surveys. We know that first time respondents are usually customers with something very good or very bad to say. Initial non-respondents are in the category of "merely satisfied" and usually have responses that are lower than the first-time respondents. By doing a parallel analysis of the initial non-respondents, we can estimate the rating of these attributes to a wider customer base. This is a way to confirm the conclusions we drew from the first-time respondents.

There were 139 surveys from initial non-respondents; and in each attribute they rated the company lower than first-time respondents. However, this group did not give us evidence that they thought price was an issue at the .05 significance level. The results show that for initial non-respondents, although price is still rated lower in customer satisfaction, the difference is not statistically significant when compared to the other attributes in contributing to customer satisfaction.

The calculations show that for initial non-respondents, the value of W corrected at the .05 significance level and 8 degrees of freedom is,

$$W = 2.95$$

Where the test disproves equality of the means anytime the value of the test statistic is greater than 15.51. In this case, initial non-respondents do not feel that price is an issue.

This conclusion is not obvious by looking at the average satisfaction scores. The formal analysis must be performed to evaluate this data.

See Figure 3 for a graph of the satisfaction scores comparing initial respondents to initial non-respondents.

Our conclusion is that price is a motivator for some of the customers, but not for all.

4. RECOMMENDATIONS

Now that we know the statistical results, we must rely on the management team to construct a root cause analysis of the reasons for this price objection. The fact that first-time respondents and first-time non-respondents have different feelings as to the company's performance in the price category is a complex problem, and it points out the importance of making that extra effort to have first-time non-respondents fill out surveys and mail them in. They represent the majority of the customers and they often have different opinions about the company. In other words, they represent two different populations; and each one needs to be analyzed separately.

5. CLOSURE

The process employed is the same one we can use for any study of customer satisfaction metrics. In summary:

- Ask customers what attributes are most important to them.

- Poll customers with current experience about the company's performance in these customer-identified attributes either continuously, or periodically (such as once a year).

- Perform a Kruskal-Wallis test on the customer satisfaction metrics for all the attributes to determine if any show statistically significant differences.

- Check the consistency of those conclusions by including first-time non-respondents (representative of the remaining customer base) in the analysis.

- Present the results to the management team for development of strategic initiatives to deal with any statistically significant differences between attributes, within the context of the business plan and the customer's perception of customer satisfaction.

This process can be employed with any quality improvement program; and it offers the confidence of having a statistical basis to discriminate between attributes that need attention and those that will not result in measurable increases in customer satisfaction even if we improved them.

To be clear on this point, any attribute that is given appropriate attention should result in improvement. However, we may maximize our use of resources by concentrating on the attributes that are the lowest, and are those that offer the most return from our investments.

Low attributes are often considered “deal-breakers” by customers, and they will not do business with companies that have very poor performance in an important attribute.

We often think we know what our customers want, but unless we ask them, we are never really sure. We often think we know what attributes our customers want us to improve, but if we work on the wrong ones, usually the ones that are easiest for us to affect, we miss the opportunity to have a measurable, positive effect with our customers.

Reference

Business Statistics in Practice, Third Edition, Bowerman, O’Connell, McGraw Hill, 2003

Sheldon D. Goldstein, P.E. is a consultant at The Steele Group, and a professor in the College of Business at Indiana Institute of Technology in Indianapolis, IN. Mr. Goldstein has a Masters degree in Mechanical Engineering from the Massachusetts Institute of Technology and an MBA in Finance from Fairleigh Dickinson University. He is a Senior Member of ASQ, and an ASQ Certified Quality Auditor.

Chi-squared Probability Density Function

Figure 1

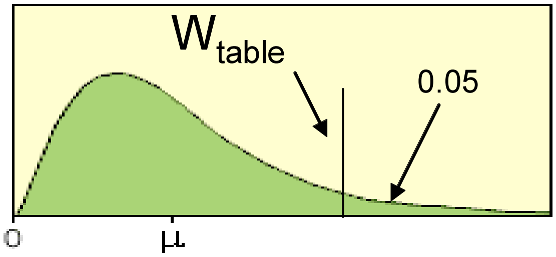


Figure 2

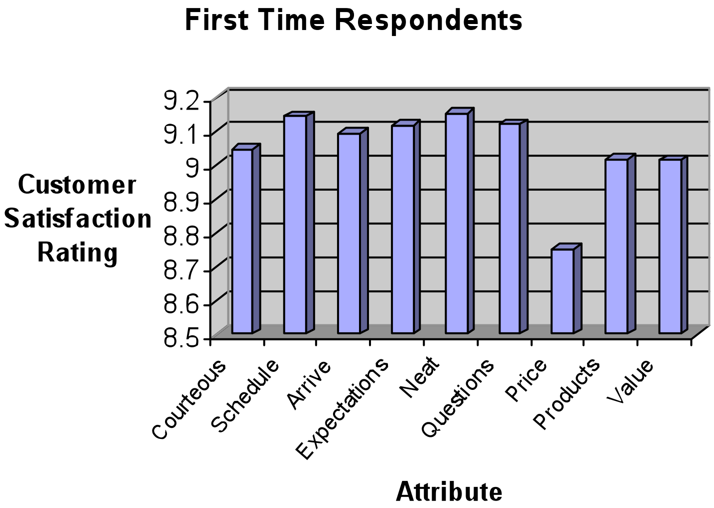
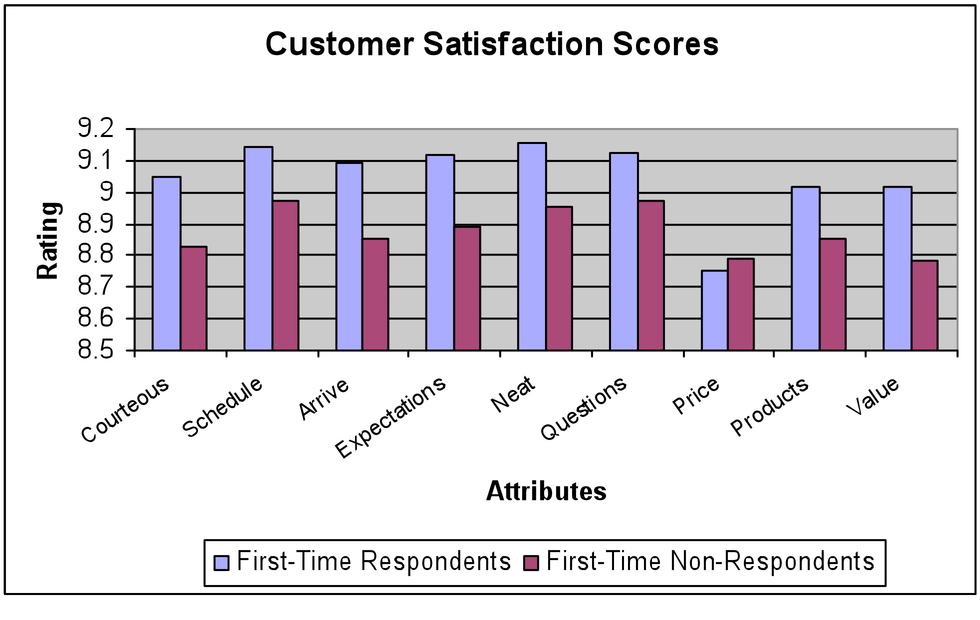


Figure 3


Customer Satisfaction Scores
Rating
Courteous
Schedule
Arrive
Expectations
Neat
Questions
Price
Products
Value
Attributes
□ First-Time Respondents
■ First-Time Non-Respondents