Presented at the 2013 ICEAA Professional Development & Training Workshop - www.iceaaonline.com

Evaluating Cost Relationships with Nonparametric Statistics

Caleb Fleming

Abstract

Parametric statistics are oft used to analyze data, evaluate hypotheses, and determine the significance of a given set of inputs. Most commonly, cost analysts use the following parametric testing measures for significance: Pearson correlation, Z-tests, T-tests, and Analysis of Variance (ANOVA) tests.

While each of these tests is subjectively appropriate to evaluate varying scenarios, each test also objectively depends on the validity of a fixed set of assumptions. Though the assumptions differ by test, all root themselves in the general idea that the data is contained within an adequately large sample, which is uncommon in federal/government acquisitions, and follows a particular probability distribution with estimable parameters.

When these assumptions are met, parametric tests are relatively simple to calculate and generate conclusions and estimates with high levels of power and accuracy. However, if the assumptions are incorrect or nonnumeric data is presented, the aforementioned tests yield invalid results viewed with skepticism. The accuracy of a parametric statistical test ultimately depends entirely on the validity of a specified set of underlying assumptions.

Cost estimators rarely operate within the parameters of a best-case-scenario, especially with regard to data acquisition. Practically, the data obtained frequently contains substantial gaps that do not follow given distributions with estimable parameters. Cost estimates are not exclusively quantitative. In these instances, the estimator can look to nonparametric statistics to evaluate relationships and determine significance.

Nonparametric statistics are useful for studying nominal and ordinal data not following a particular distribution. Nonparametric tests offer results considered robust, though less powerful, meaning there is a higher probability the test will reject a false null hypothesis. Even still, nonparametric tests may present realistic alternatives to estimating exclusively using parametric tests.

The nonparametric tests and measures addressed in this paper include: Spearman correlation, Mann-Whitney test, Wilcoxon Signed-Rank (SR) test, and nonparametric Chi-Square tests. These tests vary in purpose from identifying a correlation coefficient of two related but nonlinear variables, to testing the difference between existing and hypothetical samples that follow different distributions.

This paper examines the deficiencies in parametric tests while emphasizing the advantages of nonparametric statistics for cost estimating and the aforementioned nonparametric tests. This paper provides practical examples where a nonparametric test can determine statistical significance otherwise regarded as invalid by parametric testing.

Evaluation Cost Relationships with Nonparametric Statistics

Statistical analysis is a cardinal component of cost estimating. When utilized appropriately, tests for significance allow analysts to validate and generate Cost Estimating Relationships (CERs) that prove highly beneficial for developing strong cost estimates for a variety of programs.

Tests for statistical significance are divided into two factions -- nonparametric and parametric. Parametric estimating, renowned for its precision and usefulness in identifying distributions, is frequently favored to its nonparametric counterpart. Parametric estimating is more widely used because it is easier to understand and easier to interpret. Parametric distributions are typically normal; they yield high statistical power, and often produce accurate and precise estimates. Parametric statistics make significant assumptions about sets of data and in doing so generate specific results on what is taking place with said data points.

Unfortunately, as this paper details, the compulsory underlying assumptions for parametric estimating are frequently not met. In these instances, it is imperative for analysts and statisticians alike to understand nonparametric statistics.

This paper's purpose is to provide readers with an overview of nonparametric estimating. There are numerous nonparametric and parametric tests not detailed in the paper, and though those unmentioned tests are equally useful in determining whether observations in a dataset reflect patterns or chance, the following four measures are the subject of substantial amounts of literature and are arguably amongst the most commonly used.

Background – Statistical Principles

In order to understand nonparametric statistics and their applicability to cost estimating, a variety of statistical principles must first be introduced and defined.

Regarding statistical data, there are four fundamental levels of measurement used to describe a dataset. These levels are as follows:

- Nominal
- Ordinal
- Interval
- Ratio

These four descriptors can be further grouped into quantitative and qualitative forms of measurements. Further still, they can be considered continuous or discrete data points. In short, qualitative data is descriptive, whereas quantitative data is numerical. Discrete data points are finite in the sense that they can only take on particular values, i.e. there is no gray area between points. Continuous data points are infinite and therefore not restricted to clearly separate values.

Nominal data is exclusively categorically discrete. This type of data is simply a name, label, or category that is unable to be ordered and therefore not analyzed for standard deviation or mean. An example of nominal data includes different military services – Marine Corps, Army, Navy, etc. Though this set describes a particular type of data, specific values or rankings are not able to be determined.

Ordinal data is often numeric, though sometimes categorical, with data points measurable on an arbitrary scale capable only of distinguishing between ranks. Values presented on the ordinal scale are used to create classes and often appear in nonparametric testing. Ordinal data is simply data in which the order matters, but not the difference between the values. Examples of ordinal data include military ranks, as they indicate a general order.

Much like ordinal data, interval data represents strength in values and is ranked on a scale differentiating numerically between points. Interval data points are addressable as greater than, less than, and equal to one another. In interval datasets, zero represents another number on the scale, not the absence of a unit. An example of this type of data is temperature. On the interval scale, one temperature is not considered better than another, but rather a different number on said scale. The difference between 40 degrees and 30 degrees is the same as the difference between 70 degrees and 60 degrees.

The ratio level of measurement allows for the comparison of values by a ratio. In ratio data, zero does imply the absence of a unit. Sums and differences are calculable with ratio data. Examples of ratio data include heights, weights, and ages.

Typically, nominal and ordinal datasets should be analyzed with nonparametric statistics, and interval and ratio data should be analyzed with parametric statistics. While means, standard deviations, and standard errors of the mean can be calculated for ratio and interval data, they cannot be computed for nominal and ordinal data. Hence, in the absence of these descriptors, there is a need for nonparametric statistics to generate conclusions about specific data points.

In order to evaluate relationships and draw conclusions about the discussed levels of measurement, statistical procedures are necessary. Statistical procedures can be divided into two levels of classification– descriptive statistics and inferential statistics.[1] Descriptive statistics describe data without drawing conclusions, whereas inferential statistics involve hypothesis testing and ultimately drawing conclusions. Cost estimators commonly use descriptive statistical measures, such as the mean, range, standard deviation, variance, etc. Inferential statistics are less commonly used, but advantageous for generating conclusions in the presence of the appropriate data.

Though particular inferential hypothesis tests are addressed in greater detail below, several test components are necessary to preemptively note.

Table 1 includes the names and definitions of the components that together comprise inferential hypothesis test components.

Table 1: Inferential Test Components

Test Component	Description
Null Hypothesis (H_0)	Asserts a statement about population statistic
Alternative Hypothesis (H_1)	Asserts a claim to be tested
Significance Level (α)	Maximum probability of rejecting H_0 when H_0 is actually true
Test Statistic (T)	Value is computed to assess the validity of the null hypothesis
P-Value (p)	Probability of obtaining a test statistic at least as extreme as the one observed; The null hypothesis is "rejected" when the p-value is less than the predetermined significance level
Conclusion	(Fail to) reject H_0; There is (in)sufficient statistically significant

[1] http://sociology.about.com/od/Statistics/a/Descriptive-inferential-statistics.htm

	evidence to suggest the originally asserted statement about the population statistic.

Parametric and Nonparametric Inferential Tests

As a rule, parametric inferential tests are preferred to their nonparametric counterparts. Parametric tests are much more commonly used in the analysis of data, evaluation of hypotheses, and determination of the significance of given sets of inputs. Most commonly, cost estimators use the following parametric testing measures: Pearson's product-moment correlation, z-tests, and t-tests.

Differences between z-tests and t-tests include:

- Z-tests follow normal distributions and are generally accompanied by steeper slopes than t-distributions
- Z-tests are appropriate for datasets with larger sample sizes (30 or more)
- Z-tests are preferred when standard deviations are known

As previously mentioned, each of the parametric tests depends highly on the validity of a fixed set of assumptions. The assumptions for several tests are outlined throughout the paper, though all are rooted in the idea that the data undergoing analysis follows a particular probability distribution with estimable parameters.

In the classroom, estimators find that parametric tests are straightforward and lead to conclusions and estimates with high levels of accuracy. However, reality indicates that the assumptions are often violated. Furthermore, generating conclusions with the presence of ordinal data becomes near impossible in the parametric realm.

Nonparametric testing, however, become useful in studying ordinal datasets that do not follow normal distributions. These tests are inherently more robust, and though less powerful, present realistic alternatives to traditional parametric cost estimating procedures.

From an analytical perspective, nonparametric tests are advantageous because the probability results they generate are not constrained by the same assumptions as parametric tests. Regardless of the distribution from which a random sample was drawn, nonparametric tests offer functionality. The nonparametric realm analyzes datasets that are ranked, as well as data sets with numerical scores that have strength in ranks.

Unfortunately, nonparametric statistics notoriously fall short in power, precision, and higher ordered interactions. Many software packages are also not equipped to perform nonparametric tests. Nonparametric statistics are tools that tell the analyst that something is happening, as opposed to parametric statistics which inform the analyst of exactly what is happening.

The nonparametric tests discussed in this paper, along with their parametric counterparts, are listed in Table 2.

Table 2: Parametric vs. Nonparametric Test Comparison

Parametric	Nonparametric Equivalent
Pearson's correlation coefficient	Spearman's rank correlation coefficient
Two-sample t-test	Mann-Whitney test

Parametric	Nonparametric Equivalent
Paired t-test	Wilcoxon Signed-Rank test
Z-test	Chi-square difference in proportions

In order to understand the usefulness and application of each nonparametric test, the parametric equivalent must also be analyzed. Subsequent sections of this document pertain directly to each of the above nonparametric tests, with each section containing a brief overview of the parametric equivalent and associated assumptions. The calculation portion of each tests exclusively address nonparametric estimating. Where appropriate, examples directly applicable to cost estimating are included.

For simplicity, each of the hypotheses presented is calculated as a two-tailed test, vice lower-tailed or upper-tailed. One-tailed tests are useful for determining if the mean in a particular year for a given data set is less than or greater than in the previous year. Two tailed tests inform the statistician of whether or not the means or medians differ between years, regardless of if this difference is greater or less.

Correlation Coefficients

The first statistical measure addressed is not used for hypothesis testing, but rather to determine measures of correlation. Bivariate relationships are quintessential components of parametric and nonparametric statistics, and are generally agreed to be best measured by Pearson's correlation coefficient and Spearman's correlation coefficient, respectively.

In both the parametric and nonparametric realms, these coefficients are used to indicate the strength of the relationship between variables. The results for Pearson's coefficient range from -1 to 1, with a result of -1 indicating a perfectly negative correlation and a result of 1 indicating perfectly positive correlation.

Figure 1 shows several samples of Pearson's correlation coefficients.

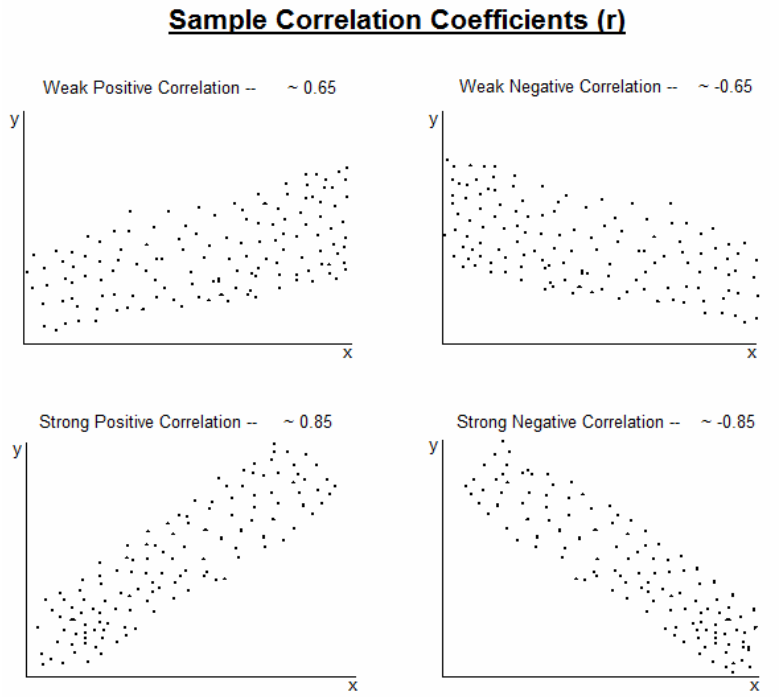

Figure 1: Pearson's Correlation Coefficient Samples

Pearson's coefficient does not differentiate between dependent and independent variables, but rather offers the analyst a rudimentary presentation of how correlated two variables are.

In the event that the data set presents normal, linear data, Pearson's coefficient is highly effective and a useful measure of correlation. Pearson's coefficient struggles, however, in areas where linearity is not present. Pearson's correlation is not robust, and its value is misleading in the presence of outliers. These outliers are observable through scatter plots.

Spearman's Correlation Coefficient

The nonparametric counterpart to Pearson's coefficient is Spearman's coefficient. This alternative coefficient is useful in detecting trends that are nonlinear but still monotonic. Monotonic trends imply that a given dataset adheres to a strictly increasing order of all ranks assigned to the parameters, though this order does not necessary need to be linear. An example of both a monotonic and non-monotonic trend is presented in Figure 2. Note that the monotonic trend displayed is not necessarily linear.

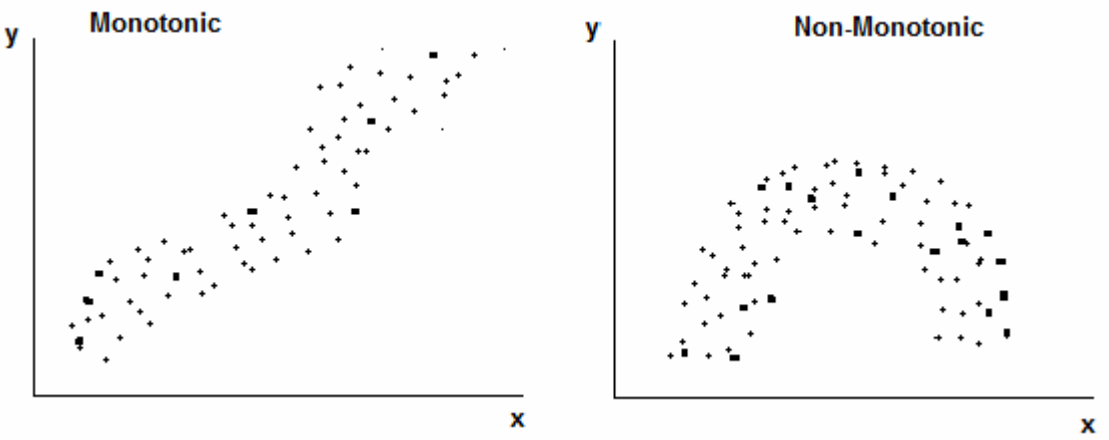

Figure 2: Spearman's Correlation Coefficient Samples

Monotonicity is an important assumption under Spearman's correlation because it indicates the existence of a relationship between the two variables, even though such a relationship is nonlinear.[2]

An example of the usefulness of Spearman's correlation is in the preferences of two test reviewers. If, for example, the performance of a particular vehicle in testing is gauged and scored subjectively by two judges, a statistician could not use Pearson's coefficient because the existence of a linear relationship between the two scores is inconsequential. Spearman's coefficient could, however, indicate if the judges agree with each other's views as far as vehicle performance goes. The test could indicate that one judge values a particular parameter more highly than others.

An additional example of the usefulness of Spearman's correlation is in fuel efficiency data. If, for example, the linear relationship between the distance traveled and quantity of fuel replenished is to be tested in the potential presence multiple outliers, Pearson's coefficient could be inaccurate and misleading. Spearman's coefficient is comparatively robust to outliers. Table 3 contains sample fuel efficiency data used to demonstrate the benefits of Spearman's coefficient in this instance.

Table 3: Fuel Efficiency Data

Distance Traveled (miles) (X)	Fuel Replenished (gal) (Y)
600	8.28
87	9.6
33	6.7
73	9.1
64	8.51
55	8
53	7.75
82	9.4
47	7.62
400	7.3
71	8.72
53	7.75

[2] https://statistics.laerd.com/statistical-guides/spearmans-rank-order-correlation-statistical-guide.php

The dependent variable (Y) in this set is the number of gallons of fuel replenished into the vehicle. The independent variable (X) is the number of miles traveled between refueling. In order to develop Spearman's coefficient, each of these variables must be ranked. These ranks are presented in Table 4.

Table 4: Ranked Fuel Efficiency Data

X	Y	Rank Mileage (X_i)	Rank Fuel QTY (Y_i)
600	8.28	12	7
87	9.6	10	12
33	6.7	1	1
73	9.1	8	10
64	8.51	7	8
55	8	5	6
53	7.8	4	5
82	9.4	9	11
47	7.62	2	3
400	7.3	11	2
71	8.72	6	9
52	7.75	3	4

The next step is to calculate the differences (D_i) between each respective X (X_i) and Y (Y_i) value, as well as the squared differences (D_i^2) between each pair ranking. This calculation is as follows:

$$D_i = X_i - Y_i$$

The calculated differences are presented in Table 5.

Table 5: Ranked Fuel Efficiency Data Differences

X	Y	X_i	Y_i	D_i	D_i^2
600	8.28	12	7	5	25
87	9.6	10	12	-2	4
33	6.7	1	1	0	0
73	9.1	8	10	-2	4
64	8.51	7	8	-1	1
55	8	5	6	-1	1
53	7.8	4	5	-1	1
82	9.4	9	11	-2	4
47	7.62	2	3	-1	1
400	7.3	11	2	9	81
71	8.72	6	9	-3	9
52	7.75	3	4	1	131

In addition to these differences between ranks, the calculation requires the differences between each rank and the mean rank for the variable. These summary statistics and differences are presented in Table 6 and Table 7. Obvious outliers are highlighted yellow, though they are included in the calculation to demonstrate the usefulness of Spearman's Coefficient in comparison with Pearson's Coefficient.

Table 6: Fuel Efficiency Descriptive Statistics

Summary	Value
Sample Size (n)	12

Mean of X Ranks	6.5
Mean of Y Ranks	6.5

Table 7: Comprehensive Fuel Efficiency Ranked Data

X	Y	X_i	Y_i	D_i	D_i^2	$X_i-\bar{x}$	$X_i-\bar{x}^2$	$Y_i-\bar{y}$	$Y_i-\bar{y}^2$	$X_i-\bar{x} * Y_i-\bar{y}$
600	8.28	12	7	5	25	5.5	30.25	0.5	0.25	2.75
87	9.6	10	12	-2	4	3.5	12.25	5.5	30.25	19.25
33	6.7	1	1	0	0	-5.5	30.25	-5.5	30.25	30.25
73	9.1	8	10	-2	4	1.5	2.25	3.5	12.25	5.25
64	8.51	7	8	-1	1	0.5	0.25	1.5	2.25	0.75
55	8	5	6	-1	1	-1.5	2.25	-0.5	0.25	0.75
53	7.8	4	5	-1	1	-2.5	6.25	-1.5	2.25	3.75
82	9.4	9	11	-2	4	2.5	6.25	4.5	20.25	11.25
47	7.62	2	3	-1	1	-4.5	20.25	-3.5	12.25	15.75
400	7.3	11	2	9	81	4.5	20.25	-4.5	20.25	-20.25
71	8.72	6	9	-3	9	-0.5	0.25	2.5	6.25	-1.25
52	7.75	3	4	1	131	-3.5	130.75	2.5	136.75	68.25

With this table developed, Spearman's coefficient, presented as rho (ρ), is calculated using the following equation:

$$\rho = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}$$

This particular dataset yields a coefficient of ρ = 0.50. Had Pearson's coefficient, presented as r, been utilized, the outliers would have skewed the data and produced a result of r = 0.008. The contrast between these two coefficients is staggering. Spearman's coefficient, without requiring for parameters to be determined to remove an outlier, present a usable coefficient that accurately predicts the relationship between the distances traveled per gallon of fuel.

Correlation Coefficients Interpreted

As with Pearson's coefficient, Spearman's coefficient is measured between -1 and 1. A coefficient less than 0 implies a negative agreement between the two variables measured. A coefficient greater than 0 implies a positive agreement between the ranks. A coefficient of 0 simply implies no agreement.[3]

Summarily, Pearson's coefficient best detects linear trends, whereas Spearman's coefficient best detects general monotonic trends. The correlation coefficient is oft regarded as one of the most significant and substantial measures in statistical analysis. As such, presenting and explaining the usefulness of such an alternative correlation measure in lieu of linearly related variables or in the presence of ranked data is valuable.

[3] http://explorable.com/spearman-rank-correlation-coefficient

Mann-Whitney Test

The Mann-Whitney nonparametric test is comparable to the two-sample t-test in parametric statistics.

The two-sample t-test is useful for determining whether or not the means of two populations on a particular outcome are equivalent. A parametric example of this would be in medical cost estimating where it is necessary to determine if the results of a controlled experiment apply to males in the same way they do females.

The nonparametric equivalent to the two-sample t-test is the Mann-Whitney test and is useful under similar circumstances. The Mann-Whitney test is useful for the comparison of samples from two groups. Statisticians use this test to determine if two distribution functions are alike, or if the expected value of one population is greater than the expected value of the other population. For the analysis of three groups, a Kruskal-Wallis (nonparametric) test for comparison is necessary.

Mann-Whitney Test Assumptions

As with parametric tests, several assumptions exist in nonparametric tests. However, the assumption requirements in nonparametric statistics are less stringent and therefore do allow for more robust estimating at the compromise of precision. Table 8 details a comparison of the assumptions for the two-sample t-test and Mann-Whitney test.[4]

Table 8: Mann-Whitney Assumption Comparison

Two-sample t-test	Mann-Whitney test
Random sampling	Random sampling
Normal distribution	Level of measurement is ordinal or interval
Equal variances	Independence within and between samples
Independence between samples	

The key difference between these two sets of assumptions, and the most appropriate indicator of when the analyst should choose nonparametric estimating over parametric estimating, is the presence of a normal distribution. As a general rule of thumb, the Mann-Whitney test is more powerful when the data does not follow a normal distribution, whereas the two-sample t-test is more powerful when the data does follow such a normal distribution.

[4] http://www.monarchlab.org/Lab/Research/Stats/2SampleT.aspx

Mann-Whitney Test Calculation

Null Hypothesis **Alternative Hypothesis**	H_o: E(X) = E(Y) H_1: E(X) ≠ E(Y)
Level of Significance	α = 0.05
Observed Test Statistic	$Z = \frac{U - E(U)}{\sqrt{\frac{N_1 N_2 (N+1)}{12}}}$
Variables Defined	Z= z score U = Sum of rank scores for given set E(U) = Expected rank score N = Total sample size (Inclusive of both data sets) N_1 = Sample size of data set 1 N_2 = Sample size of data set 2
P-value	2 x Min[P(α ≤ 0.05), P(α ≥ Z)]
Conclusion	If the P-value above is less than, we would reject the null hypothesis and conclude there is statistically significant evidence to suggest that E(X) ≠ E(Y)
Critical (Test) Statistic	Obtained from z-table at the significance level α

Example:

The Mann-Whitney test is applicable in a situation in which an estimator must discern whether or not there is a statistically significant difference in quality between two widgets produced by differing manufacturers. Such a sample dataset is included in Table 9.[5]

Table 9: Widget Manufacturing Data

Widget Number	Manufacturer	Widget Score
1	1	221
2	1	245
3	1	269
4	1	205
5	1	175
6	1	229
7	1	237
8	1	261
9	1	148
10	1	186
11	1	237
12	1	202
13	1	261
14	2	218
15	2	194
16	2	240
17	2	232

[5] http://www.isixsigma.com/tools-templates/hypothesis-testing/making-sense-mann-whitney-test-median-comparison/

Widget Number	Manufacturer	Widget Score
18	2	229
19	2	267
20	2	127
21	2	213
22	2	237
23	2	269
24	2	205
25	2	223

The hypothesis to be tested in this example is that there is not statistically significant evidence to prove that either location, on average, produces higher rated widgets. A rejection of this hypothesis would indicate a relationship between the manufacturing location and the widget score, whereas a failure to reject said hypothesis would confirm the suspicions.

As with most nonparametric tests, the first step in the calculation is to aggregate and rank the data. Instead of ranking by each respective subset, the Mann Whitney test joins the data sets and ranks cumulatively across the different samples. This is demonstrated in the table low. For the sake of calculating the sample sizes necessary for the test statistic calculation, the cumulative rankings are displayed in Table 10 and further broken out.

Table 10: Ranked Widget Manufacturing Data

Widget Number	Manufacturer	Widget Score	Combined Ranks	Location 1	Location 2
1	1	221	11	11	-
2	1	245	20	20	-
3	1	269	24	24	-
4	1	205	7	7	-
5	1	175	3	3	-
6	1	229	13	13	-
7	1	237	16	16	-
8	1	261	21	21	-
9	1	148	2	2	-
10	1	186	4	4	-
11	1	237	16	16	-
12	1	202	6	6	-
13	1	261	21	21	-
14	2	218	10	-	10
15	2	194	5	-	5
16	2	240	19	-	19
17	2	232	15	-	15
18	2	229	13	-	13
19	2	267	23	-	23
20	2	127	1	-	1
21	2	213	9	-	9
22	2	237	16	-	16
23	2	269	24	-	24
24	2	205	7	-	7
25	2	223	12	-	12

Upon obtaining the rankings, the analyst must generate sums of ranks for each individual location, as well as for the comprehensive set. These statistics are displayed in Table 11.

Table 11: Widget Manufacturing Ranked Sums

Category	Rank Sum	Sample Size
Cumulative	318	25
Location 1	164	13
Location 2	154	12

With these tables compiled, the test statistic can now be calculated with the following equation:

$$E(U) = \frac{N_1(N+1)}{2}$$

In this particular example, the value obtained using the first location is 169, and the value obtained using the second location is 156. The lesser of the two values should be inserted into the following equation to get a Z-score.

$$Z = \frac{U - E(U)}{\sqrt{\frac{N_1 N_2 (N+1)}{12}}}$$

The Z-score obtained using this equation is 4.24, and the p-value associated with this score for a two-tailed probability is 0 to the thousandth digit. Based on our earlier hypothesis test, this result would lead us to reject our null hypothesis at a 95% level of confidence and conclude there is a statistically significant difference between the scores of the two widget manufacturers. Further, because the rank sum for the first location is greater than that of location two, it is safe to assume that on average, location two produces a higher rated product.

Wilcoxon Signed-Rank (SR) Test

The Wilcoxon Signed-Rank nonparametric test is comparable to the paired t-test.

From a parametric standpoint, the paired t-test is useful when the analyst is presented a single measurement variable and two nominal variables. For example, this test would be useful for measuring the fuel efficiency of a ground vehicle before and after receiving an upgrade to its engine. In this instance, the measurement variable would be the fuel efficiency, and the nominal variables would be a particular vehicle's respective before and after performance rating. This test is considered more reliable when the difference between groups is small relative to the variation within groups. [6]

The nonparametric equivalent – the Wilcoxon SR test – is used when the normality assumption is violated or the sample size is too small and a need remains to measure paired data points. In the aforementioned example, normality could be violated in the presence of outliers. Often times the analyst is uncertain as to whether or not outliers are present. Rather than arbitrarily removing data, the Wilcoxon SR test could be used. For example, the user of a particular vehicle may have entered an extra digit or omitted an existing digit when entering mileage. The Wilcoxon SR test is robust to handle such outliers in a way that the paired t-test would not.

Wilcoxon SR Test Assumptions

Table 12 details a comparison of the assumptions for the paired t-test and Wilcoxon SR test.

Table 12: Wilcoxon SR Test Assumption Comparison

Paired t-test	Wilcoxon SR test
Random sampling	Random Sampling
Normal distribution	Continuous dependent variable
Equal variances	Level of measurement is at least ordinal
Dependence within sample pairs, independence between samples	Dependence within sample pairs, independence between samples

Much like in the Mann-Whitney test, the normality assumption is a key indicator of which test should be chosen. If all of the parametric assumptions are met, the paired t-test will result in a precise, robust estimate. However, if the normality assumption is violated, the Wilcoxon SR test offers a robust and useful alternative.

[6] http://udel.edu/~mcdonald/statpaired.html

Wilcoxon SR Test Calculation

Null Hypothesis **Alternative Hypothesis**	H_o: E(X) = E(Y) H_1: E(X) ≠ E(Y)
Level of Significance	α = 0.05
Observed Test Statistic	If n ≤ 50: $T = \sum_{i=1}^{n} R_i$ If > 50: $T = \frac{\sum_{i=1}^{n} R_i}{\sqrt{\sum_{i=1}^{n} R_i^2}}$ R_i = Sum of sign-ranks T = Test statistic
P-value	2 x Min[P(T_2 ≤ Z), P(T_2 ≥ Z)]
Conclusion	If the P-value above is less than α, we would reject the null hypothesis and conclude there is statistically significant evidence to suggest that E(X) ≠ E(Y)
Critical (Test) Statistic	Obtained from t-table using N degrees of freedom (DOF) at the significance level α

In the Wilcoxon SR test, two random-sample datasets are joined so that each data point within each set is affiliated with a coordinate pair. Each of these bivariate pairs is then converted back to univariate data by finding the difference and absolute difference within the pair. Observations with an absolute difference equal to zero are removed and not included in the sample size.

The remaining observations are ranked from smallest to largest by absolute difference value and assigned sign-ranks. These absolute ranks are then assigned sign-ranks based on the original difference between the two coordinate pairs. For example, if a paired rank is 8, and the original difference was -3, their sign-rank would simply equal -8.

The sum of the sign-ranks is then used as the test statistic and compared with its associated critical value. If the sample size is greater than 50, an additional calculation is required to generate such a test statistic. Additionally, as the sample size increases, the Wilcoxon SR distribution begins to converge to a normal distribution.

Example:

The Wilcoxon SR test could be applied to a situation in which an analyst desires to compare the functionality provided by two masks exposed to the presence of chemical agents. The data set in Table 13 contains 11 points -- or cases. This sample size would likely be regarded as inadequate for parametric testing. Each case has an associated number of hours that the mask was able to withstand the chemical agent before becoming susceptible to leakage.

Table 13: Chemical Agent Resistance Capabilities Data

Case	Mask 1 (Hours)	Mask 2 (Hours)
1	7.2	10.5
2	9.9	17.1
3	8.1	10.8
4	7.5	6.3
5	22.8	29.1
6	11.4	10.5
7	42.6	18.6
8	18.3	9.6
9	7.8	12.3
10	8.1	30.9
11	17.7	63

The hypothesis tested in this example is that there is no statistically significant evidence to prove that one mask outperforms the other. A rejection of our null hypothesis would imply that the durability ratings of the masks are effectively equal, while a failure to reject the null hypothesis would indicate the opposite.

As previously stated, each of the masks has been assigned a coordinate pair, simply by the way the data was collected. It is assumed that the masks were exposed to the same chemical agent at the same time, and that one proved itself more durable by preventing leakage for the specified number of hours. The next step in conducting this test is to find the absolute value of the differences between the coordinate pairs. Following this, the differences will be ranked and assigned a positive or negative sign, depending on what the original difference value was. The results are contained within Table 14.

Table 14: Ranked Chemical Agent Resistance Capabilities Data

Case	Mask 1 (Hours)	Mask 2 (Hours)	Actual Differences	Absolute Values	Ranked Differences	Sign-Rank
1	7.2	10.5	-3.3	3.3	4	-4
2	9.9	17.1	-7.2	7.2	7	-7
3	8.1	10.8	-2.7	2.7	3	-3
4	7.5	6.3	1.2	1.2	2	+2
5	22.8	29.1	-6.3	6.3	6	-6
6	11.4	10.5	0.9	0.9	1	+1
7	42.6	18.6	24	24	10	+10
8	18.3	9.6	8.7	8.7	8	+8
9	7.8	12.3	-4.5	4.5	5	-5
10	8.1	30.9	-22.8	22.8	9	-9
11	17.7	63	-45.3	45.3	11	-11

The next step in this test is to sum the total positive ranks and the total negative ranks. In this instance, with a relatively small dataset, this is a simple task. The sum of positive ranks is 21, and the sum of negative ranks is 45. Combining these two, we find that the cumulative rank sum is -24. By taking the absolute value of this integer, the test statistic is obtained.

In order to generate a conclusion with this test statistic, a number of degrees of freedom (DOF) must be generated. DOF are the number of values in a calculation that are variable,[7] typically measured as the number of pairs less one. In this example, there are 10 DOF.

Observing this test statistic and value for DOF in the context of the critical values table for the Wilcoxon SR, a critical value of 8 is realized. Because our test statistic of 24 is significantly larger than the critical value of 8, the null hypothesis is rejected and it is determinable that statistically significant evidence exists to suggest that the protection offered by the two masks is not equal.

[7] http://www.animatedsoftware.com/statglos/sgdegree.htm

C**hi-Square Difference in Proportions Test**

The chi-square difference in proportions test is similar to an Analysis of Variance (ANOVA) parametric test. In these tests the response variables are dichotomized into classes.

ANOVA tests are used to determine statistically significant differences between multiple groups. Conclusions are drawn in ANOVA testing by the ratio of two variances which are not affected by constant bias or scaling errors. ANOVA tests are preferred to a standard t-test, which becomes unreliable when more than two samples are included.[8] Unfortunately, multiple pairs of proportions dictate the necessity of a chi-square test.[9]

A chi-square distribution is derived from the normal distribution, and is essentially composed of a sum of squared z-scores. Figure 3 represents a two-sided Chi-Square Probability Density Function (PDF) Distribution.[10]

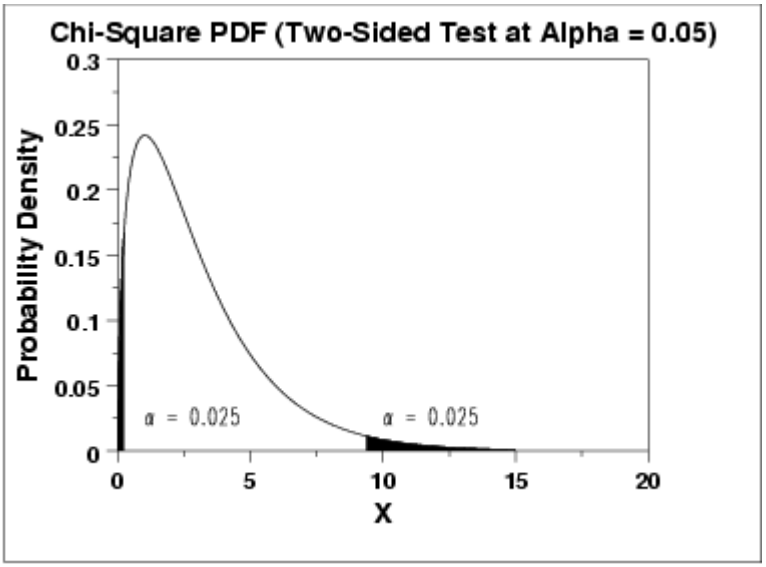

Figure 3: Chi-Square Distribution Sample

Each independent term is a DOF and the DOF, paired with the desired significance level, make up the chi-square probability table. The chi square distribution compares expected frequencies of occurrence within contingency tables to their observed frequencies. If the observed and expected frequencies are too far apart, the null hypothesis is rejected.

Particularly, the chi square test for proportion compares the equality of two or more population proportions. Previously, tests for independence between variables have been studied.[11]

These tests are particularly useful for fitting statistical models to observed data. If an analyst desires to determine how near observed values are to what would be expected under the guidelines of the fitted model, the chi-square difference in proportions test would apply.

[8] http://explorable.com/anova
[9] http://facstaff.unca.edu/dohse/online/stat185e/unit5/st5_2_chisq_l.htm
[10] http://www.itl.nist.gov/div898/handbook/eda/section3/gif/chspdftb.gif
[11] http://www.stat.wmich.edu/s216/book/node115.html

Chi-Square Difference in Proportions Test Assumptions

Table 15 details the assumptions for the chi-square difference in proportions test.

Table 15: Chi-Square Difference in Proportions Assumption Comparison

ANOVA Test	Chi-Square Difference in Proportions
Normality	Independence between samples
Independence between samples	Each observation occurs once
Homoscedasticity	Adequate sample size

Chi-square tests excel in investigating distributions and proportions of categorical variables by using tallies. Chi-square is also beneficial in determining if the observed values correspond with the expected values. Chi-square tests are robust.

Chi-Squared Difference in Proportions Test Calculation

Null Hypothesis **Alternative Hypothesis**	H_o: $P_{1j} = P_{2j} = \ldots = P_{rj}$, for all j H_a: $P_{ij} \neq P_{kj}$, for some i, j, k
Level of Significance	$\alpha = 0.05$
Observed Test Statistic	$T = t = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij} - E_{ij})^2}{E_{ij}}$ df = (n-1)
P-value	$p_v = 2(\min p_1, p_2)$ $p_1 = P(T_1 \leq t)$ $p_2 = P(T_1 > t)$
Conclusion	If the P-value above is less than α, we would reject the null hypothesis and conclude there is statistically significant evidence to suggest that $P_{1j} = P_{2j} = \ldots = P_{rj}$, for all j
Critical (Test) Statistic	Obtained from chi-square table using N DOF at the significance level α

Chi-square tests utilize contingency tables to the compare observed values and expected values. These contingency tables are then surveyed to generate a test statistic. By combining the sample proportions from multiple groups, an estimate is generated that provides more information than any of the sample proportions could provide individually.[12]

Example:

The chi-square test for proportions could be applied in a situation in which a military service desires to understand what type of attacks it is most vulnerable to in order to plan armament upgrades. The following sample dataset includes 228 observations. These observations include incidents in which a vehicle was destroyed via an improvised explosive device (IED), rocket propelled grenade (RPG), small-arms fire, or roll-over event. In each of the instances, survival

[12] http://www.prenhall.com/behindthebook/0136149901/pdf/Levine_CH12.pdf

statistics are reported for the attending crew members. The 228 figure represents soldiers, not vehicles. For example, if a single vehicle containing two soldiers was attacked, the number reported would include two of the 228 observations, not one. The raw data is compiled in Table 16.

Table 16: Vehicle Accident Data

Table of Raw Counts			
Cause/Outcome	**Fatal**	**Nonfatal**	**Total**
IED	26	31	**57**
RPG	47	16	**63**
Rollover	25	47	**72**
Small Arms Fire	14	22	**36**
Total	**112**	**116**	**228**

The hypothesis to be tested in this example is that there is no statistically significant evidence to suggest that the cause of a particular vehicle accident influences the survival outcome likelihood of the vehicle's passengers.

Figure 4 presents the above raw data as a comparison between the cause of the accident and the outcome realized.

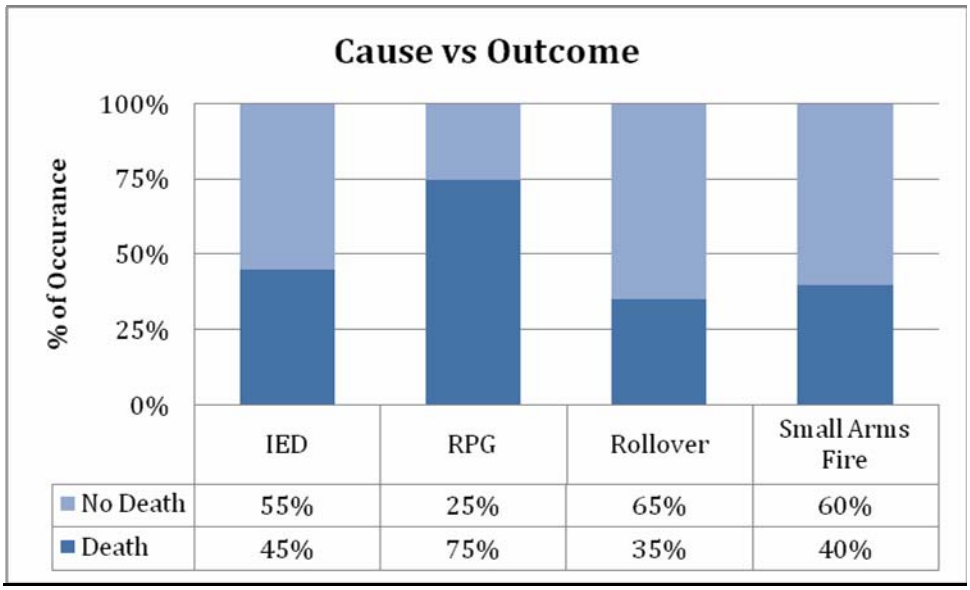

Figure 4: Cause vs Outcome Mosaic Chart

The raw data table presented above serves as the table of observed values. In order to generate a Chi-Square test statistic, an equivalent table representing expected values must also be generated. The expected values are calculated by multiplying each cell's associated row total by its associated column total, and then dividing that figure by the cumulative total. For this example, the expected value of IED deaths is calculated by multiplying 57 and 112, and then dividing by 228 to obtain a value of 28. The calculated expected values are presented in Table 17.

Table 17: Expected Outcomes by Cause

Table of Expected Counts

Table of Expected Counts			
Cause/Outcome	**Fatal**	**Nonfatal**	**Total**
IED	28	29	**57**
RPG	30.95	32.05	**63**
Rollover	35.37	36.63	**72**
Small Arms Fire	17.68	18.32	**36**
Total	**112**	**116**	**228**

With these tables calculated, the chi-square sums can be calculated by squaring the difference between the observed and expected values and dividing by the expected values. For our example, the chi-square sum for the IED deaths is calculated by subtracting 28 from 26 and dividing by 28. These sums are presented in Table 18.

Table 18: Chi-Square Sums (Calculated from Observed and Expected)

Chi-Square Sums			
Cause/Outcome	**Fatal**	**Nonfatal**	**Total**
IED	0.14	0.14	**0.28**
RPG	8.33	8.04	**16.37**
Rollover	3.04	2.93	**5.97**
Small Arms Fire	0.77	0.74	**1.501**
Total	**12.28**	**11.85**	**24.13**

The cumulative sum of each of the individual chi-square sums yields a value of 24.13. This value, paired with the dataset's DOF [(# of rows – 1) x (# of columns – 1)] is compared with the critical value obtained from the chi-square table (11.143).

The test statistic in our example is significantly larger than the critical value, therefore the null hypothesis is rejected and the conclusion is drawn that there is statistically significant evidence to suggest the cause of a particular vehicle accident influences the survival outcome likelihood of the vehicle's passengers.

As a follow-on analysis, each of these causes could be compared individually, such that a conclusion could ultimately be drawn to determine which of the four incidents is most lethal.

Concluding Remarks

There are numerous nonparametric tests omitted from this paper. The purpose of this document, however, is not to serve as a nonparametric cost estimating guide, but rather to present an alternative lens through which data analysis can be viewed.

While the aforementioned tests present differing methods for analyzing data and observing relationships between variables, there is still no single test for statistical significance that strictly dominates all the others. Different statistical methods, both parametric and nonparametric, provide analysts with differing perspectives of datasets and offer general insights that may otherwise have gone unnoticed.

Each of the aforementioned tests, parametric and nonparametric, have positive and negative attributes. Nonparametric statistics, while requiring less stringent and rigid assumptions, are less powerful because of their reliance on the ordinal positioning of pairs. However, in cases where this ordinal or nominal data is present, nonparametric statistics could potentially offer a solution not seen in the parametric realm.

Contrarily, in the presence of interval or ratio levels of measurement, parametric statistics provide powerful results and utilize means and standard deviations to generate realistic estimates of correlation. Parametric tests, however, require more information in their development and are more widely understood among cost estimators.

Choosing wisely between nonparametric tests and parametric tests is extremely important. From a general standpoint, not all distributions are normal, not all of the assumptions mentioned above are met, and not all of the data is received in quantitative packages. By possessing the knowledge and skill-set to perceive data from the nonparametric realm as well as the parametric realm, analysts are able to present clients with a completely alternative perspective that could lead to a better understanding of component functionality, point estimates with life cycle cost estimates, risk bounding, and general cost estimating relationships used in the development of factors.

Acronyms

Acronym	Definition	Page
ANOVA	Analysis of Variance	2
CER	Cost Estimating Relationship	3
DoD	Department of Defense	3
DOF	Degrees of Freedom	17
H_0	Null Hypothesis	5
H_1	Alternative Hypothesis	5
IED	Improvised Explosive Device	22
JPEO-CBD	Joint Program Executive Office - Chemical and Biological Defense	3
MCSC	Marine Corps Systems Command	3
p	Probability	6
ρ	Rho	11
Ri	Sum of Signed Ranks	17
RPG	Rocket Propelled Grenade	22
SAS	Statistical Analysis System	26
SR	Signed Rank	16
T	Test Statistic	6

References/Relevant Links*

*Additional materials regarding particular tests are available upon request, including the Statistical Analysis System (SAS) code for each of the tests studied.

The following resources were helpful in developing the content of this paper and are highly useful in forming a deeper understanding of the topic at hand. Nonparametric statistics, much like parametric statistics, are able to be studied at a much higher level of detail than what is presented in this document. I strongly encourage readers to peruse the contents of each of the below websites and documents to develop a more complete understanding of statistics, particularly in the nonparametric realm.

A significant amount of detail in this paper is developed using information obtained through lecture slides developed by Marlow Lemons, an Advanced Instructor in the Department of Statistics at Virginia Tech. Additionally, the tables outlining each of the components of the hypothesis tests, were developed using

General Statistics/Nonparametric Statistics Overview:

Nonparametric Statistics Academic Coursework: Lemons, Marlow. "Stat 3504 Nonparametric Statistics." Virginia Tech Fall 2010. Lecture.
Nonparametric Statistics Application Paper: Fleming, Caleb & Vaughn, Samuel. "Population Demographics and Crime Distributions." Undergraduate coursework. Virginia Tech, 2011.
Levels of Measurement: http://statistics.about.com/od/HelpandTutorials/a/Levels-Of-Measurement.html
Descriptive and Inferential Statistics: http://sociology.about.com/od/Statistics/a/Descriptive-inferential-statistics.htm
Degrees of Freedom: http://www.animatedsoftware.com/statglos/sgdegree.htm
SAS coding: http://support.sas.com/publishing/pubcat/chaps/62097.pdf
Differentiating Nonparametric and Parametric Realms:
http://www.csse.monash.edu.au/~smarkham/resources/param.htm

Spearman's Correlation Coefficient vs. Pearson's Correlation Coefficient:
Defining Pearson's Coefficient: http://www.statisticshowto.com/articles/what-is-the-pearson-correlation-coefficient/
Defining Spearman's Coefficient:
http://www.johnmyleswhite.com/notebook/2009/02/17/pearson-vs-spearman- correlation-coefficients/ & http://explorable.com/spearman-rank-correlation-coefficient *Comparing Pearson's and Spearman's Coefficients:* https://statistics.laerd.com/statistical-guides/spearmans-rank-order-correlation-statistical-guide.php

Mann-Whitney Test
Definition and Calculation Overview: http://www.unm.edu/~marcusj/WMW.pdf
Hypothesis Testing: http://www.isixsigma.com/tools-templates/hypothesis-testing/making-sense-mann-whitney-test-median-comparison/
Comparison to Two-Sample T-Test:
http://www.monarchlab.org/Lab/Research/Stats/2SampleT.aspx

ANOVA Test
Definition: http://en.wikipedia.org/wiki/Analysis_of_variance#Characteristics_of_ANOVA
Test explanation: http://explorable.com/anova

Wilcoxon Sign-Rank Test
Definition: http://en.wikipedia.org/wiki/Wilcoxon_signed-rank_test
Practical Example:
http://www.utdallas.edu/~serfling/3332/Practical_Guide_Wilcoxon_Signed_Rank_Test.pdf
Step-by-Step Directions: http://www.vassarstats.net/wilcoxon.html &
http://mlsc.lboro.ac.uk/resources/statistics/wsrt.pdf
Comparison to Paired T-Test: http://udel.edu/~mcdonald/statpaired.html

Chi-Square Test for Proportions
Definition and Test Explanations:
http://www.prenhall.com/behindthebook/0136149901/pdf/Levine_CH12.pdf ;
http://facstaff.unca.edu/dohse/online/stat185e/unit5/st5_2_chisq_l.html ;
http://www.stat.wmich.edu/s216/book/node115.html
Distribution Sample:
http://www.itl.nist.gov/div898/handbook/eda/section3/gif/chspdftb.gif
SAS Discussion:
http://people.stat.sfu.ca/~cschwarz/Stat-650/Notes/PDFbigbook-SAS/SAS-part015.pdf

Probability Tables

Mann-Whitney Probability Table
http://www.lesn.appstate.edu/olson/stat_directory/Statistical%20procedures/Mann_Whitney%20U%20Test/Mann-Whitney%20Table.pdf

Wilcoxon Sign Rank Probability Table
http://facultyweb.berry.edu/vbissonnette/tables/wilcox_t.pdf

Chi-Square Probability Table
http://econ.clarion.edu/econ222/eagle222/chisquaretable.htm

Standard Normal (Z) Probability Table
http://www.doe.virginia.gov/testing/test_administration/ancilliary_materials/mathematics/2009/2009_sol_z_table.pdf

Biography

Caleb Fleming
Kalman & Company, Inc.
2101 Wilson Boulevard, Arlington, VA, 22201
(571) 388-5767; caleb.fleming@kalmancoinc.com

Caleb Fleming is a cost analyst supporting various DoD clients. Mr. Fleming is well-versed in the development of life-cycle cost analysis, including the collection and analyses of data, generation of estimate assumptions and methodologies, and creation of cost estimating relationships (CERs). Mr. Fleming is experienced in utilizing parametric and nonparametric statistics in analyses and has functionally applied such concepts -- including those addressed in the subsequent paper -- to aid in the development of life-cycle cost estimates (LCCEs). Mr. Fleming has supported numerous clients within the Marine Corps Systems Command (MCSC) and the Joint Program Executive Office for Chemical and Biological Defense (JPEO-CBD). His programmatic experience extends from tactical wheeled vehicles to counter-fire radars and chemical and biological detection devices. Mr. Fleming was awarded a bachelors degree in Economics from Virginia Tech in 2011. Additionally, he minored in Statistics and was the winner of several national writing awards.