2011 NDIA GROUND VEHICLE SYSTEMS ENGINEERING AND TECHNOLOGY SYMPOSIUM
MODELING & SIMULATION, TESTING AND VALIDATION (MSTV) MINI-SYMPOSIUM
AUGUST 9-11 DEARBORN, MICHIGAN

HYPOTHESIS TESTING AND COMPARISON OF FUEL CONSUMPTION TEST RESULTS

Todd A. Morris
Automotive Instrumentation Division
U.S. Army Aberdeen Test Center
Aberdeen Proving Ground, MD 21005

ABSTRACT

Fuel use and resupply provisions are vital to all military combat operations, and any significant reduction in the amount of fuel required to sustain the force becomes a large tactical advantage for commanders Developers have been seeking methods that offer even small gains in fuel economy. Small gains for a fleet of thousands of vehicles translate into fewer fuel convoys to theater and large costs savings over time. The challenge to the tester and evaluator is to determine if these small advances are relevant or merely normally expected test variation in the acquired fuel consumption parameters. All too often only the mean fuel economy parameters are compared with and without the new equipment or process without considering test variances inherent in collecting the parametric data. The resulting analysis may then be seriously flawed. Hypothesis testing is a useful statistical method for comparing two sets of test data (sample means and standard deviations) to determine if there is a statistically significant difference between the two sets. Often the two sets of data are made up of small sample sizes (5 test trials are very typical for sets of fuel consumption data). Therefore, for purposes of this discussion, we will consider only hypothesis tests for the differences between two sample means for small (the number of samples is less than 30) sample sizes. Several examples of fictional test data will be subjected to hypothesis testing to show the value of such an approach.

INTRODUCTION

It's well known that our current military operations are heavily dependent on fuels, especially those burned in generators, aircraft and vehicle systems. Fuel shipments to current war a zones in a recent article in National Defense Magazine are on the order of 60 to 70 million gallons per month[1]. In 2009, Pentagon officials speaking before the House Appropriations Committee cited that the average fully burdened fuel cost (the cost to acquire, transport and protect fuel) to supply fuel to remote outposts in Afghanistan was $400 per gallon[2]. It's also an accepted fact that fuel convoys in current theaters of operation are prized targets of opportunity for insurgents. According to an article in the Washington Post, a recent Marine Corps study found that one Marine is wounded for every 50 trips made for fuel or water in Afghanistan[3]. All defense departments have been focused to find savings in the amount of fuel used and alternatives to traditional fuels. Any significant reduction in the amount of fuel required to sustain the force becomes a large tactical advantage for commanders can mean fewer fuel convoys and significant cost savings.

Very often in fuel consumption testing, customers are seeking an improvement in the amount of fuel consumed by a vehicle system. Developers have been seeking methods that offer even small gains in fuel economy. Small gains for a fleet of thousands of vehicles translate into fewer fuel convoys to theater and large costs savings over time. The challenge to the tester and evaluator is to determine if these small advances are relevant or merely normally expected test variations in the acquired fuel consumption parameters. All too often in my experience only the mean fuel economy parameters are compared with and without the new equipment or process without considering test variances inherent in collecting the parametric data. Often test funding does not allow for a large number of samples for each parameter. The resulting analysis may then be seriously flawed.

ANALYSIS ISSUES

A series of examples may illustrate the potential for analysis flaws. We'll consider two fictitious populations – men born in Norway in 1955, and men born in the United States in the same year. As a variable, we'll pick the current height of adult men from these two populations, and we'll assume that someone spent a good bit of time measuring each member's height simultaneously. You'll remember from your grade school days that if a variable is normally distributed, the distribution of the variable in the population will resemble a bell shaped curve centered on the population mean for the assessed variable. Figure 1 shows the two distributions of the population heights.

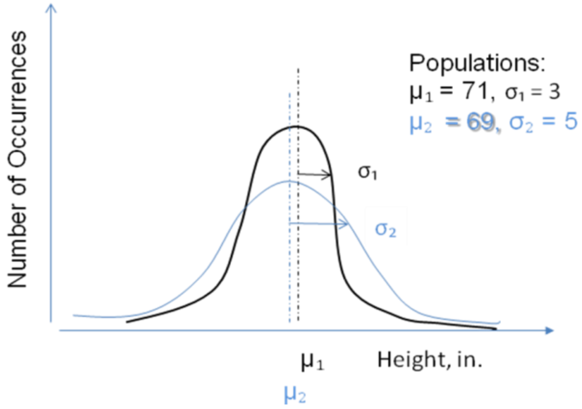

Figure 1:Height distributions of two different populations.

As you can see, the two population distributions for height seem to indicate bell-shaped curves, and the population means (μ_1 and μ_2) indicate that for our fictitious populations, the mean height of men from Norway is greater than the mean height of men from the United States. Probably more important to an analysis is the relative shapes of these two distributions. A general rule of thumb for a normally distributed random variable states that 68 percent of the area under the distribution curve will lie within one standard deviation of the mean value of that parameter.[4] The wider and flatter distribution for U.S. men indicates that there are a wider variety of heights for that population, and the standard deviation of the height, σ_2, will be larger as a result. The Norwegian population is more uniform with respect to height, and the standard deviation, σ_1 is smaller.

The problem for testers and evaluators is that we hardly ever get to sample every member of our test populations. We sample from that population. For a fleet of thousands of vehicles to be fielded, we may only get one example. Budgets for testing may be limited, and as a result, we may only be able to run three to five iterations of a test scenario with that single vehicle. Instead of dealing with population means (μ), we're typically dealing with a five sample average,$\bar{x}$, and a sample standard deviation, s. So, for our example with the heights of populations, say we sent someone to Norway to take the heights of ten randomly selected men, and we did the same here in the United States. The data may look like Table 1.

TABLE 1. SAMPLE HEIGHT DATA FOR TWO FICTITIOUS POPULATIONS

Sample	Height, Norwegian adult male, in.	Height, U.S. adult male, in.
1	72	73
2	69	72
3	68	68
4	65	70
5	73	70
6	66	68
7	70	65
8	69	72
9	68	71
10	67	71
Average, $\bar{x}$	69	70

Considering only the sample averages, you would conclude (incorrectly for our fictitious populations) that adult males in the United States are taller than their Norwegian counterparts. By rolling the dice and taking only a small sample size for a large population, and then compounding the problem by considering only the difference between your sample means without considering the impact of your sample variance, you've set yourself up for a high potential for analysis error. Comparison of means alone is a very common practice, and in some cases it is appropriate. But is there a way to compare data with small sample sizes to determine if meaningful differences do exist? Fortunately, the answer is affirmative.

HYPOTHESIS TESTING FOR TESTS WITH SMALL SAMPLE SIZES - BACKGROUND

Hypothesis testing is a useful method to compare two sets of test data to determine if there is a statistically significant difference between the two sets. Often the two sets of data are made up of small sample sizes due to limited budgets and a limited number of prototype test vehicles (5 test trials for a single vehicle are very typical for sets of fuel consumption data that we collect at our test center). Therefore, for purposes of this discussion, we will consider only hypothesis tests for the differences between two sample means for small (the number of samples is less than 30) sample sizes.

Because we deal with small sample sizes, we cannot assume that the test statistic for fuel consumption will be a normally distributed random variable, and it is more appropriate to use the Student's t-distribution instead of the normal distribution.[4] In many respects, the Student's t-distribution is similar to the normal distribution, in that the distribution is bell shaped, but is much more appropriate for small sample sizes. The Student's t-distribution has an interesting history, and we actually have the Guinness Brewery to thank for it. A man named William Gosset first developed the ideas for the t-distribution working with samples of the raw materials for the making of beer, but Guinness did not allow its employees to publish scientific papers. The reasons for this ban aren't exactly clear, but Gosset saw great potential for his work, and published his initial work under the name Student.[5] Subsequent follow-on work by a number of other mathematicians fleshed out the theory, but the name stuck.

In order to use hypothesis testing and the Student's t-distribution, we must assume that the population of fuel consumption data from which we are sampling from has an approximately normal probability distribution, and that the samples are selected independently (in other words, we acquire the data without changing the way we conduct the tests).

Hypothesis testing relies on establishment of a sample mean and standard deviation from the acquired fuel consumption data, and relies on the evaluator to provide an expected level of confidence in the results. Typically, the level of confidence is established at 95% or 99%, which means that once the hypothesis test is conducted and the statistics are calculated, we are 95% or 99% confident that our statistical test provides a correct decision. These levels of confidence are closely tied to the area under the probability density function (pdf). In general terms the ability to render a decision on comparisons of sample means between two populations increases as the area of overlap of the two distributions decreases. Figure 2 depicts this phenomenon graphically. In the first group of figure 2, our populations have similar means and there is a high degree of overlap between the two probability density functions. We would have very little confidence that there is a statistically significant difference between samples from these two populations. The second portion of the figure shows similar probability density functions, with less overlap. Less overlap means less shared area under each of the density functions, and we would have more confidence that there is a statistically significant difference between the two population means. The last figure shows even more separation, less area of overlap, and hence increasing confidence. One of the easiest ways to accomplish the final portion of Figure 2 is to decrease variability (standard deviation) of the sample data.

.

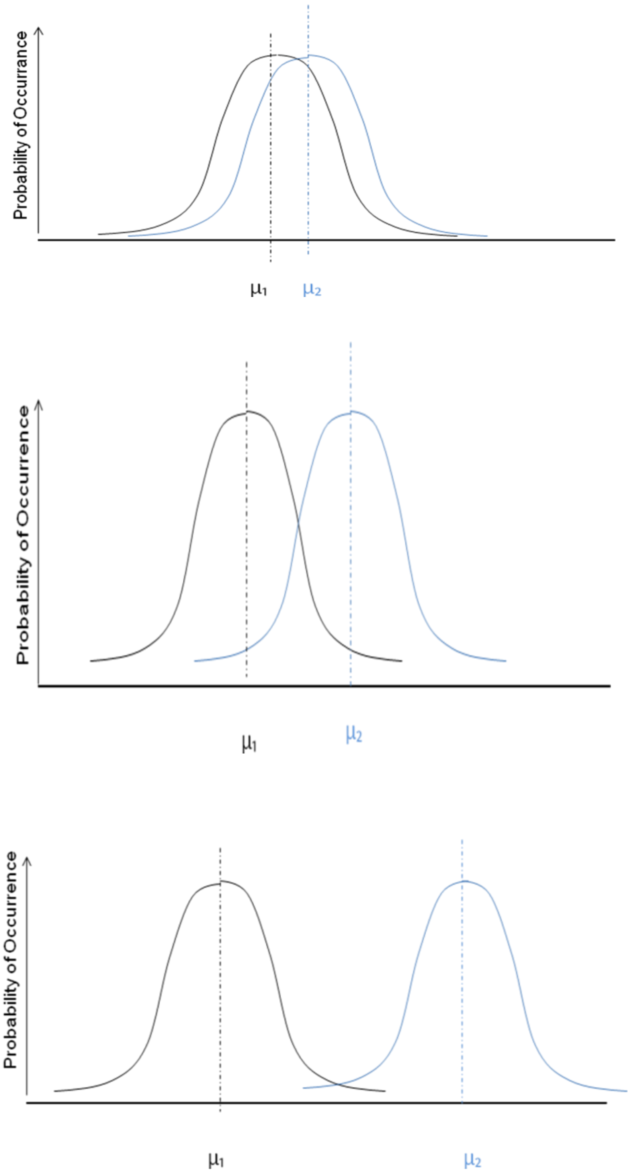

Figure 2: Three examples of probability density functions with varying mean differences.

HYPOTHESIS TESTING EXAMPLES AND IMPLEMENTATION

For purposes of explanation, consider the following fictional example for fuel consumption test data. A materiel developer wants to apply a fuel additive to the fuel system of a 5-ton truck and would like to know if the additive provides a statistically significant increase in fuel economy. The first set of five fuel consumption tests were run with the standard fuel mixture, and the second set were run in the same way with the fuel additive. Test results are presented in Table 2.

TABLE 2. SAMPLE FUEL CONSUMPTION DATA FOR A FICTITIOUS TEST PROGRAM

Sample	Without Additive, Pounds per Hour	With Additive, Pounds per Hour
1	39.6	38.3
2	41.1	40.6
3	38.4	39.1
4	39.7	38.8
5	40.5	40.0
Sample mean, $\bar{x}$	39.9	39.4
Standard Deviation, s	1.02	0.93

Let's compare the means and standard deviations for these two sets of data. Without the additive, the mean is 39.9 pph, while the standard deviation is 1.02 pph. With the additive, the mean is 39.4 pph, with a standard deviation of 0.93 pph. By comparing only the means of the two data sets, one might argue that the additive decreased the amount of fuel burned by 0.5 pph. Thousands of 5-ton trucks burning fuel in theater on a daily basis means that even a 0.5 pph difference can probably be sold as a decrease in costs. But is this difference merely statistical chance?

We'll conduct a hypothesis test to determine if the difference is statistically significant. In this case, we must pick an assumption (a null hypothesis in statistical terms), and we'll assume that for purposes of our test that the two means of the fuel consumption populations are equal – that there is no difference between the means. Alternatively, we'll pick an alternative hypothesis that the mean of our fuel consumption population obtained with the fuel additive is less than that obtained without the fuel additive. Because we're indicating that in our alternate hypothesis that one population will be better than the other, we're conducting what's called a one-tail hypothesis test. The other option (which won't be discussed in this example) would be a two-tailed hypothesis test, in which we'd propose as an alternative hypothesis that the fuel consumption would either be better or worse than the fuel consumption obtained with the standard fuel mixture. For this example, a one-tailed hypothesis test for fuel consumption parameters will be conducted at a one-percent level of significance (or a 99% level of confidence) using the following formulas:[4]

$$H_0 : \mu_0 = \mu_1,$$

$$H_1 : \mu_0 > \mu_1,$$

$$t = \frac{\bar{x}_0 - \bar{x}_1}{s_p\sqrt{\frac{1}{n_0}+\frac{1}{n_1}}},$$

$$\text{and, } s_p = \sqrt{\frac{(n_0-1)s_0^2+(n_1-1)s_1^2}{n_0+n_1-2}},$$

where H_0 is the null hypothesis (the two population means for fuel usage are equal), H_1 is the alternate hypothesis (the population mean for fuel usage without the additive is greater than the fuel usage with additive), μ_0 is the population mean for fuel usage without the additive, μ_1 is the population mean for fuel usage with the additive, *t* is the test statistic to be compared with the student-t distribution, $\bar{x}_o$ is the mean of the fuel use samples without the additive, $\bar{x}_1$ is the mean of the fuel use samples with the additive, n_0 is the number of samples taken without the additive, n_1 is the number of samples taken with the additive, s_0 is the standard deviation of the fuel use samples taken without the additive, and s_1 is the standard deviation of the fuel use samples taken with the additive.

Using the fuel consumption values, means and standard deviations listed above, we substitute into the equations to calculate the t-statistic.

$$s_p = \sqrt{\frac{(5-1)(1.02)^2+(5-1)(0.93)^2}{5+5-2}} = 0.976$$

$$t = \frac{39.9-39.4}{0.976\sqrt{\frac{1}{5}+\frac{1}{5}}} = 0.81$$

The test statistic, *t*, for this test scenario is then compared to the Student's t-distribution statistic at a one-percent level of significance, $t_{0.010}$ using $n_0 + n_1 - 2$ degrees of freedom. Student's t-distribution tables are available in statistical textbooks as charts listed by degrees of freedom and levels of significance. A 1-percent level of significance means we have a 1-percent chance of rejecting a null hypothesis when we should accept it, and we have a 1-percent chance of accepting a null hypothesis when we should reject it. For the test scenario (five trials without the additive ($n_0 = 5$) and five trials with the additive ($n_1 = 5$)), there would be 8 degrees of freedom. Thus, by consulting a table of the Student's t-distribution, the Student's t-statistic for rejection of the null hypothesis (that the two sample means are equal) is established at 2.896. The null hypothesis would be rejected (the fuel usage with the additive was significantly less than the baseline condition) only if *t* is greater than 2.896. In our case, t was 0.81, so we cannot reject the null hypothesis that the fuel consumption with and without the additive is any different. We would then state that we cannot say that the additive had a statistically significant effect on the fuel consumption.

In general, the ability to reject the null hypothesis and determine that two populations are statistically different is aided by increasing the sample size and decreasing the variability of the test statistic. Tests should be designed (if funding is available) with large numbers of samples, and variability in test methods should be minimized.

For example, what if we were able to secure funding to do two more test runs with and without the additive, and we only tested the vehicle with the same driver for each test to reduce possible differences in driving through the test course. Maybe our data might look Table 3.

TABLE 3. SAMPLE FUEL CONSUMPTION DATA FOR A FICTITIOUS TEST PROGRAM

Sample	Without Additive, Pounds per Hour	With Additive, Pounds per Hour
1	40.0	39.2
2	40.1	39.2
3	39.8	39.1
4	39.7	39.2
5	40.1	40.0
6	39.8	39.5
7	39.6	39.5
Sample mean, $\bar{x}$	39.9	39.4
Standard Deviation, s	0.20	0.31

By comparing only the means, we would again think that the additive decreased the amount of fuel burned by 0.5 pph. We should also note that our standard deviations are now smaller. Let's consider our hypothesis test and accompanying statistics.

Using the fuel consumption values, means and standard deviations listed above, we substitute into the equations to calculate the t-statistic.

$$s_p = \sqrt{\frac{(7-1)(0.2)^2+(7-1)(0.31)^2}{7+7-2}} = 0.263$$

$$t = \frac{39.9-39.4}{0.263\sqrt{\frac{1}{7}+\frac{1}{7}}} = 3.56$$

The test statistic, *t*, for this test scenario is then compared to the Student's t-distribution statistic at a one-percent level of significance, $t_{0.010}$ using $n_0 + n_1 - 2$ degrees of freedom. For the test scenario (seven trials without the additive ($n_0 = 7$) and seven trials with the additive ($n_1 = 7$)), there would be 12 degrees of freedom. Thus, by consulting a table of the Student's t-distribution, the Student's t-statistic for rejection of the null hypothesis (that the two sample means are equal) is established at 2.681. The null hypothesis would be rejected (the fuel usage with the additive was significantly less than the baseline condition) only if *t* is greater than 2.681. In our case, t is 3.56, so we can reject the null hypothesis, and accept the alternate hypothesis that the fuel additive decreases fuel consumption. You can see from this example, that increasing the sample size and decreasing variability in the results can make a small difference in means statistically significant.

CONCLUSION

We can see that the application of hypothesis testing can be quite powerful and convincing. There are very few occasions in the real world when we'll be able to say we're confident in a result, and be able to quantify the level of our confidence. By considering the variability of test data, and armed with some simple equations for the calculation of test statistics, hypothesis testing can be a useful tool. In terms of fuel consumption, even small improvements in fuel economy would be embraced, but there must be confidence that the means of accomplishing the improvement are truly effective.

REFERENCES

[1] S. Irwin, "Army's Energy Battle Plan: Attack Fuel Demand", National Defense Magazine, May 2011.

[2] R. Tiron, "$400 Per Gallon Gas To Drive Debate Over Cost of War in Afghanistan", The Hill, October 15, 2009.

[3] Associated Press, "Marines Going Green to Save Lives On Battlefield, Avert Attacks On Fuel Convoys In Afghanistan", Washington Post, March 21, 2011.

[4] J. Newmark, "Statistics and Probability in Modern Life", Harcourt Brace College Publishers, Orlando, 1992.

[5] S. Zabell., "On Student's 1908 Article "The Probable Error Of A Mean", Journal of the American Statistical Association, March, 2008.