



Safety Management Services, Inc.
Clint Guymon, PhD PE
1847 West 9000 South Suite 205
West Jordan, UT 84088

January 21, 2012

Subject: The Chart Significance Method and PROBIT Comparisons (Draft)

Safety Management Services, Inc. provides hazards analysis services to the energetic materials manufacturing industry. Hazards analysis includes identifying areas of the process where the in-process energies used could approach the material's ignition energies through impact, friction, electrostatic discharge (ESD), or other impetus. The energies used in the process are in most cases evaluated qualitatively and recommendations for increasing the safeguards (or not) are determined from such a qualitative comparison coupled with decades of experience. Occasionally, quantitative comparisons between the in-process energy and the initiation energy are needed to determine the ignition probability. Such occasions include accident investigations and critical processing steps.

Quantitative hazards analysis consists of sensitivity testing and in-process energy analysis. Sensitivity testing can yield the probability of initiation at a given energy level and when compared against the in-process energy yields the initiation probability at a given process step. The probability of initiation is found by subjecting the energetic material to multiple trials of friction, impact, ESD or other impetus and recording the number of reactions observed.

This article discusses the statistical inferences that can be made when comparing sensitivity testing results. In several instances there is a need to compare the results of two sensitivity test results and draw conclusions from such. A couple of examples of such situations include using sensitivity testing to determine sample differences or to determine machine, operator, or site repeatability. Discussed here are two methods to compare sensitivity testing results: Significance Chart Method and PROBIT.

Comparison between the characteristic responses (initiation probability versus impetus energy or the number of reactions in a given number of trials) of two materials is valuable for use in decision making. One material may have a different slope relating energy to reaction probability or a different number of reactions for a given number of trials but that difference may or may not be statistically significant. If the difference is not statistically significant, making conclusions to the contrary may lead to negative consequences. For example, a new formulation may appear to be more sensitive than the previous formulation but statistically it may be inconclusive whether or not it's different. Modifying the formulation could be a costly and incorrect response to such a result.

Before discussing the two methods, it's important to understand the variability with binomial trials. Binomial trials are those where the outcome is either a success (no reaction) or failure (reaction). An example is flipping a coin (in a perfectly random way). The probability of flipping a head with an unbiased coin is 50%, yet if a coin is flipped randomly ten times, the result could be anywhere from 0 heads to 10 heads, although it's likely (95% probable) that between 2 to 8 heads will be observed. It cannot be concluded that because only 2 heads in 10 trials were observed that the coin was biased. It likely would take significantly more trials than 10 to identify a biased coin.

In the discussion presented here, any variability in operating the testing equipment (such as variability due to machine inconsistencies or operator or environmental conditions) is not included. Although this undoubtedly could be a significant factor, the effect on the variability of the sensitivity testing is not addressed here other than stating that if a significant difference is observed between samples, the difference may be due to operator or equipment inconsistencies. This discussion treats the best case where the operational variability (differences between operators, reaction determination, and machine operation) is limited or insignificant.

Chart Significance Method

Perhaps the quickest way to evaluate two energetic samples to determine similarities or differences in sensitivity is to test at a given energy. At a given energy, Sample A may yield 3 of 20 reactions whereas Sample B gives 9 of 20 reactions. From these results are the two samples different? Often it's concluded that they are different. The Chart Significance Method (developed by Safety Management Services, Inc.) gives a statistical based answer to the above question. The method applies to constant energy binomial testing.

As hinted at above in the introductory section, there is a distribution of outcomes resulting from binomial testing. We have observed persons in industry make conclusions where that inherent distribution is not properly weighted. The Chart Significance Method makes it easy to successfully do so. Table 1 below makes it easy to compare the results from testing of two 20 trial samples at a given energy. Likewise Table 1 is for two 10 trial samples.

Each cell in Table 1 represents a hypothesis test between two sets of 20 trials. The color corresponds to the level of significance (by p-value) in rejecting the hypothesis that the probabilities of initiation for the two tests of 20 trials are equivalent (with the alternative hypothesis being that they are not equivalent). For example, PETN exhibits 3 reactions in 20 trials at a 16 cm drop height for impact sensitivity and Sample C exhibits 9 reactions in 20 trials at that same height. Using Table 1, it cannot be stated with statistical significance that Sample C is different than PETN. Table 1 shows the results of such a hypothesis test represented by the cell where row 3 (representing 3 reactions in 20 trials) intersects with column 9 (representing 9 reactions in 20 trials); the cell is gray indicating that it cannot be stated with 95% confidence that the initiation probabilities as tested of the two materials differ. However, if Sample C had resulted in 10 reactions in

20 trials it could have been concluded that Sample C is more sensitive than PETN with 95% confidence. A detailed description of the methodology and steps of such a hypothesis test are given below in the Methodology section of the Appendix

The variability in the observed reaction probability, e.g. 4 of 10 reactions seen followed by perhaps 7 of 10 reactions, is reduced as the number of trials increases. This analysis assumes that the probability of initiation is exactly the same each and every trial; if variation is introduced from either the operator, machine, or substance the initiation probability is likely no longer constant across trials. The results in Table 1 and 2 are the best case (i.e. the inconclusive band is as narrow as it can get) where the initiation probability is constant for each trial.

PROBIT Comparison

A PROBIT plot relates the event probability to an energetic impetus. In energetic manufacturing and testing the impetus is usually impact, friction, or ESD. PROBIT plots are useful in estimating the event probability or initiation probability at impetus values that have not been specifically tested; they are also useful in comparing the sensitivity of two conditions or materials. Comparing two material's sensitivity through PROBIT plot comparison can be more accurate than a comparison at a single energy level. A method to compare PROBIT plots is reviewed here.

PROBIT plots present non-linear behavior in a linear way. For example, most initiation phenomena are normally distributed with transition areas (regions where the probability of initiation changes from near 0 to near 1) of varying width. Figure 1 shows an example plot where at low energy the probability of initiation is low and at high energy the probability is high. PROBIT plots represent the curve linearly thus showing the low probabilities that are close to zero with better resolution.

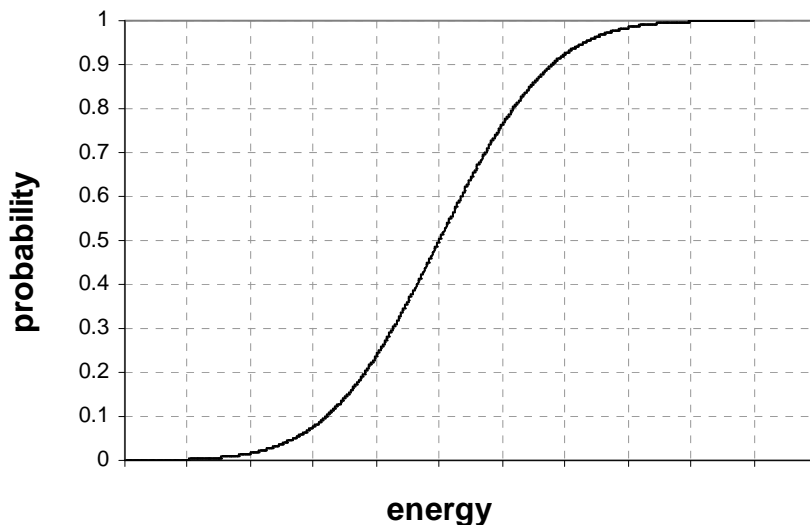


Figure 1 Typical transition from a low to a high probability of initiation as a function of impetus energy (commonly referred to as an S curve).

Comparison between the characteristic responses (initiation probability versus the impetus energy) is valuable in determining material or parametric differences. Here we describe a method to compare linearly represented initiation data. The below method is similar to the Hercules Parallel Line Assay program which was used to compare different sets of PROBIT data and to combine them to get a representative line to use for quantitative analysis.

There are simple methods that have been included in many software packages to statistically compare linear regression coefficients. The linear coefficients describing the initiation probability as a function of energy can easily be obtained when plotting the data on a PROBIT plot. The specific details used to perform a statistical comparison of the parameters of two linear relationships are described in the Appendix, here we discuss example results.

Suppose that an impact test is completed on two substances that yield PROBIT plots. The sensitivity results for the two substances are plotted in a log-normal way; experience at Hercules Inc . Aerospace Division (now part of Alliant Techsystems Inc.) indicates that a log-normal relationship best describes the relationship between the impetus energy and the probability of initiation.

A simple way to compare the two PROBIT relationships is to compare the slopes and intercepts of the regression lines. If the slopes or intercepts are statistically different then it's likely the materials have different sensitivities; however, if the slopes and intercepts are not statistically different, it cannot be concluded that the substances have statistically different sensitivities (given, of course, the log-normal relationship is true). Details of the methodology used to make such a conclusion are in the Appendix.

Conclusion

Simply comparing sensitivity differences numerically (4 of 10 versus 6 of 10 reactions) or by visual comparison (Substance B appears to be more sensitive than Substance A on the PROBIT plot) can lead to inaccurate conclusions and costly business decisions. In order to make accurate conclusions or inferences from quantitative sensitivity testing, it's imperative to use statistics. With the relatively small number of trials that are performed during sensitivity testing, statistical comparison between sensitivity results of binomial trials can be completed in a number of ways. This paper describes two such methods to statistically compare quantitative sensitivity testing results: Significance Chart Method and PROBIT.

Table 1 Matrix of significance for 2 sets of 20 binomial trials Unique p-values given are an average of two Monte-Carlo calculations of 10,000 random points each of the given distribution, modeled as a beta function. Areas in green indicate results are significant (at 95% confidence) whereas areas in grey indicate that results are inconclusive. See text for a more in depth discussion including examples. The darker hued diagonal and center square indicates regions about which the table is symmetric: table is bisymmetric (symmetric and centrosymmetric). Note that for the cases of zero reactions and 20 reactions in 20 trials, the initiation probabilities from which the table was generated are estimated; i.e. it may be possible at those levels the p-values are lower than represented and reported here.

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20			
0																							20	
1																								19
2																								18
3																								17
4	0.09																							16
5	0.05	0.13																						15
6	0.03	0.07																						14
7		0.04	0.10																					13
8			0.05	0.11																				12
9			0.03	0.07																				11
10				0.04	0.07																			10
11					0.04	0.08																		9
12						0.04	0.08																	8
13							0.05	0.08																7
14																								6
15																								5
16																								4
17																								3
18																								2
19																								1
20																								0
	20	19	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1	0			

Table 2 Matrix of significance for 2 sets of 10 binomial trials Unique p-values given are an average of two Monte-Carlo calculations of 10,000 random points each of the given distribution, modeled as a beta function. Areas in green indicate results are significant (at 95% confidence) whereas areas in grey indicate that results are inconclusive. See text for a more in depth discussion including examples. The darker hue diagonal and center square indicates regions about which the table is symmetric: table is bisymmetric (symmetric and centrosymmetric). Note that for the cases of zero reactions and 10 reactions in 10 trials, the initiation probabilities from which the table was generated are estimated; i.e. it may be possible at those levels the p-values are lower than represented and reported here.

	0	1	2	3	4	5	6	7	8	9	10	
0												10
1												9
2												8
3												7
4												6
5												5
6												4
7												3
8												2
9												1
10												0
	10	9	8	7	6	5	4	3	2	1	0	

Appendix

Methodology: Significance Chart Method

This section gives the steps taken to perform a hypothesis test on two sets of binomial trials where the sample size is small. First is described an example of a hypothesis test when comparing binomial results where the sample size is large. Then the binomial distribution and the uncertainty in a set of binomial trials are reviewed. Finally a Monte Carlo approximation is given that is used to determine the p-values of binomial proportions when the sample size is reduced.

When the number of trials is large (greater than 100 or so) and when the probability is not close to 0 or 1, hypothesis testing can be based on a normal approximation to the binomial distribution. A test statistic is computed from which conclusions on whether or not a hypothesis (e.g. the initiation probability of Sample 1 is equal to Sample 2) can be rejected in favor of an alternate hypothesis (e.g. the initiation probability of Sample 1 is not equal to Sample 2) based on the normal distribution. The test statistic is (D.C. Montgomery, G.C. Runger, and N.F. Hubele; Engineering Statistics. John Wiley & Sons, Inc., 1998.)

$$Z_0 = \frac{\frac{X_1}{n_1} - \frac{X_2}{n_2}}{\sqrt{\frac{X_1 + X_2}{n_1 + n_2} \cdot \left(1 - \frac{X_1 + X_2}{n_1 + n_2}\right) \cdot \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}, \quad (1)$$

where X_1 and X_2 are the number of reactions and n_1 and n_2 are the number of trials with the subscripts referring to the given material or circumstance. The test statistic is then compared with the cumulative standard normal distribution to determine the p-value. The p-value is “the smallest level of significance that would lead to rejection of the null hypothesis” (ibid.) where the null hypothesis in this discussion is that the initiation probability of Substance A is equal to Substance B. For example, if 100 reactions were observed in 200 trials for Substance A and 120 reactions in 200 trials for Substance B, the test statistic is equal to 2.01 yielding a p-value of 0.044; the conclusion would then be that at 95% confidence (0.044 is less than 0.05), the sensitivity of Substance A is not equal to Substance B, but at 99% confidence (0.044 is not less than 0.01) it can not be concluded that the sensitivities of the two materials are different.

Because performing hundreds of sensitivity trials is not economical in many cases, most sensitivity testing is completed with 10 to 20 trials at a given energy. For such small sample sizes, the normal approximation to the binomial distribution is no longer accurate. Prior to discussing an appropriate method to obtain the p-value for a hypothesis test where the sample sizes are small, the binomial distribution is reviewed.

Below in Figure A1 is a plot of the probability distribution showing the probability density when the number of reactions is 1 in 10 trials. The normalized distribution was

generated by varying the probability of obtaining a reaction on any given trial, p , in the binomial probability distribution with n (number of trials) equal to 10 and i (number of reactions) equal to 1. The binomial probability at a given p , i , and n is

$$P(i, n, p) = \frac{n!}{i!(n-i)!} \cdot p^i \cdot (1-p)^{n-i}. \quad (2)$$

As is evident from the plot, a wide range of initiation probabilities can result in the outcome of 1 reaction in 10 trials with the most likely scenario being that the initiation probability is 10%. More specifically, if 1 reaction in 10 trials is observed, there is a 95% likelihood that the initiation probability lies between 0.023 to 0.413.

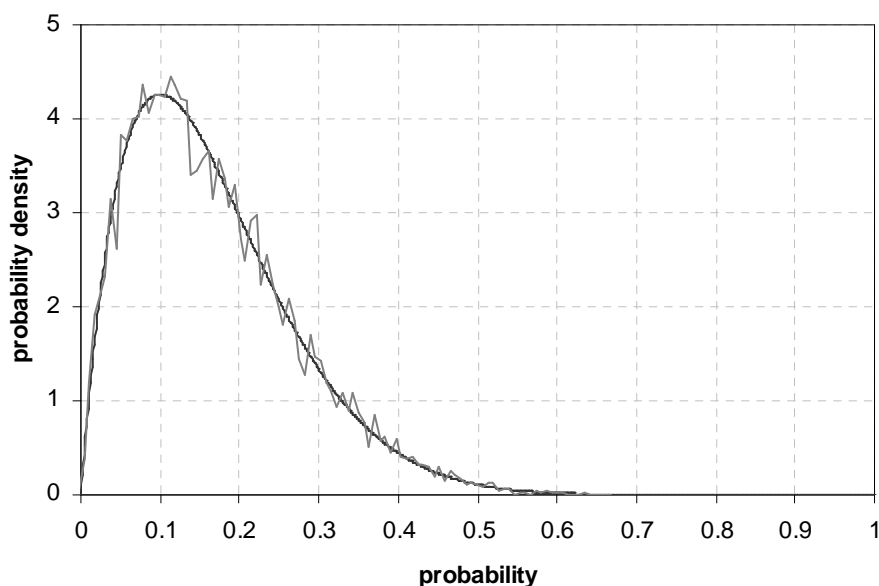


Figure A1 Probability distribution showing the probability density versus probability when the number of reactions is 1 in 10 trials. The grey line represents a normalized histogram of 10,000 points generated to approximate the original.

A Monte Carlo based method can be used in order to find the likelihood that the sensitivity of two substances are different based on small sets of binomial trials; or in other words a Monte Carlo method can be used to obtain the p-value for a hypothesis test between two sets of binomial trials where the sample sizes are small. The Monte Carlo method used here includes generating a large number of points for each data set that reproduce the respective binomial distribution for a given number of observed reactions in a given number of trials. Shown in Figure A1 is also a plot (in grey) representing a normalized histogram of 10,000 points approximating the original binomial distribution where 1 reaction was observed in 10 trials. Once the two distributions of random probabilities have been generated, the difference of the distributions is calculated from which the p-value can be found. A mathematical description of the Monte Carlo method used is given below.

The cumulative binomial distribution is approximated with a cumulative beta distribution where the alpha and beta parameters are $x+1$ and $n-x+1$, respectively where x is the number of reactions and n the number of trials. The cumulative beta distribution that yields the cumulative probability of the binomial distribution given n and x is

$$P_{cum} = I(p_{trial}, \alpha, \beta) \quad (3)$$

where P_{cum} is the cumulative probability, and p_{trial} is the trial probability.

Random numbers distributed as the binomial distribution can be generated from the inverse cumulative beta distribution. A significant number of random numbers ($\sim 10,000$) uniformly distributed are generated. Each random number is then operated on to yield the desired distribution:

$$p_{trial,k} = I^{-1}(u_k, \alpha, \beta) \quad (4)$$

where I^{-1} denotes the inverse cumulative beta distribution, u_k is a uniformly distributed random number (ranging between 0 and 1).

Once the two desired distributions, one for each set of x and n , of random numbers have been generated, the difference can be obtained. The difference between the two large numbers of random numbers can be expressed as

$$p_{diff,k} = I^{-1}(u_{1,k}, \alpha_1, \beta_1) - I^{-1}(u_{2,k}, \alpha_2, \beta_2) \quad (5)$$

where p_{diff} is termed the difference distribution and the subscript 1 and 2 refer to two independently generated distributions subject to the respective alpha and beta parameters. The difference distribution corresponds to the difference between the two binomial distributions of given x and n . If the difference distribution is centered about zero, it's likely that the null hypothesis (the initiation probability of Situation A is equal to Situation B) cannot be rejected in favor of the alternative hypothesis (initiation probabilities of Situation A and B are unequal).

Once the difference distribution is obtained, it is approximated as a Student t-distribution. A hypothesis test is completed where the null hypothesis is that the mean of the difference distribution is equal to zero and the alternative hypothesis that it is not equal to zero. The test statistic for the Student t-distribution is

$$t = \frac{x_m - 0}{s / \sqrt{n}} \quad (6)$$

where x_m is the average and s is the standard deviation of the 10,000 or so points defining the difference distribution, and n is 1 (as in this case the standard deviation of the sampling distribution is approximately equal to the standard error). The degrees of

freedom for computation of the p-value using the Student t-distribution is estimated to be $(1/n_1+1/n_2)^{-1}$, where n_1 and n_2 are the number of trials for each test condition.

Methodology: PROBIT Comparison

This section discusses the mathematical relations that can be used to statistically compare the slope and intercept of two lines. Such a comparison yields the p-values that indicate the statistical confidence of differences between the two slopes and the two intercepts. Comparing the two parameters that describe a linearized relationship is simpler than comparing parameters of a non-linear relationship.

Experience at Hercules Inc. Aerospace Division (now part of Alliant Techsystems Inc.) indicates that a log-normal relationship best describes the relationship between the impetus energy and the probability of initiation. Mathematically the relationship is

$$P(x) = \frac{1}{2} \left[1 + \operatorname{erf} \left(\frac{\log E - \log \mu_g}{\log \sigma_g \cdot \sqrt{2}} \right) \right] \quad (7)$$

where erf is the error function, P is the initiation probability, E is the impetus energy, μ_g is the geometric mean, and σ_g is the geometric standard deviation. Fitting a line to the variables $\operatorname{erf}^{-1}(2 \cdot P - 1)$, where erf^{-1} is the inverse error function, versus $\log E$ will yield the slope and intercept of the linear log-normal relationship. The slope (m) and intercept (b) relate to the geometric mean and geometric standard deviation as given in Equations 8 and 9 respectively:

$$\log \mu_g = -\sqrt{2} \cdot b \cdot \log \sigma_g \quad (8)$$

$$\log \sigma_g = \frac{1}{\sqrt{2} \cdot m} \quad (9)$$

The geometric mean corresponds to the 50% initiation probability and the geometric standard deviation can be used to approximate the confidence interval about that value. For example, if the geometric mean is 100 energy units and the geometric standard deviation is 1.1 energy units, then the energy value that lies 2 standard deviations above and below the mean is 121 ($=100 \cdot 1.1^2$) and 82.6 ($=100/1.1^2$) respectively.

In addition to determining the parameters describing the relationship between the initiation probability and the impetus energy, a comparison, or hypothesis test, between the two relationships can also be completed. As with comparing the number of reactions for two samples at a given energy, a hypothesis test is conducted to find the p-value where the null hypothesis is the slopes or intercepts are equivalent (with the alternative hypothesis that they are unequal).

The method used to complete the hypothesis test is to use dummy variables and complete a multivariable regression analysis. The method is described briefly here from SPSS

FAQ: “How can I compare regression coefficients between two groups?” UCLA: Academic Technology Services, Statistical Consulting Group. From <http://www.ats.ucla.edu/stat/spss/faq/compreg2.htm> (accessed April 10, 2009). Additional details can be found in statistical reference books such as N.R. Draper and H. Smith, Applied Regression Analysis, 3rd Edition, John Wiley and Sons, 1998.

In the multivariate regression used to perform the hypothesis test described above, the data from both sets of substances is included with the following dummy variables: dummy variable one (\mathbf{D}_1) is equivalent to 1 for data that is from Substance A and 0 for data that is from Substance B, dummy variable 2 (\mathbf{D}_2) is equal to $\log E$ for Substance A and 0 for substance B. The independent variable (\mathbf{Y}), $[\text{erf}^{-1}(2 \cdot P - 1)]$, and primary dependent variable (\mathbf{X}), $\log E$, are the same. Equation 10 shows the relationship between the variables, dummy variables, and linear parameters.

$$\mathbf{Y} = \beta_3 \cdot \mathbf{X} + \beta_2 \cdot \mathbf{D}_1 + \beta_1 \cdot \mathbf{D}_2 + \beta_0 + \varepsilon \quad (10)$$

where ε (an array) is the error between the predicted and actual values of the independent variable \mathbf{Y} , and $\beta_0 - \beta_3$ are linear coefficients. Variables in bold indicate arrays of values.

Many programs, including Microsoft Excel, can perform analysis of variance (ANOVA) to determine the level of significance for a linear regression coefficient. The specific equations and relations that are used to determine regression statistics are not quoted here. A multivariate ANOVA is performed with the above variables to determine the p-value of the hypothesis tests previously discussed comparing the two slopes and two intercepts. The coefficient of regression (β_1) for the dummy variable \mathbf{D}_1 is the difference between the intercepts found for each substance; the coefficient of regression (β_2) for the dummy variable \mathbf{D}_2 is the difference between the slopes. If either coefficient (β_1 or β_2) is significant then it's likely the materials have different sensitivities; however, if the slopes and intercepts are not statistically different, it cannot be concluded that the substances have statistically different sensitivities (given, of course, the log-normal relationship is true). Significance of the regression coefficient can be measured by calculation of the p-value; e.g. a p-value less than 0.05 indicates that the regression coefficient is significant by at least 95% confidence.