2020 NDIA GROUND VEHICLE SYSTEMS ENGINEERING AND TECHNOLOGY SYMPOSIUM
SURVIVABILITY TECHNICAL SESSION
AUGUST 11-13, 2020 - NOVI, MICHIGAN

BALLISTIC VALIDATION TEST STATISTICS AND CONFIDENCE LEVELS

James Eridon, Scott Mishler

General Dynamics Land Systems, Sterling Heights, MI

ABSTRACT

Ballistic validation testing typically involves firing multiple shots at a nominal velocity and ensuring the target stops every round with only partial penetrations, no completes. This testing is specified as a consequence of the binary nature of the test, and the need to meet a particular probability of penetration at a specified velocity with a certain confidence level. This legacy process has significant shortcomings owing to both the test procedures involved as well as the nature of the statistical interpretation of the results. This paper describes an alternative test and analysis procedure that produces the required level of performance and confidence information at a specified velocity, as well as the confidence over a wide range of other velocities and performance levels. In addition, this procedure eliminates many of the shortcomings associated with the legacy "no penetration" test protocol, and requires no more shots at the target.

Citation: J. Eridon, S. Mishler, "Ballistic Validation Test Statistics and Confidence Levels", In *Proceedings of the Ground Vehicle Systems Engineering and Technology Symposium* (GVSETS), NDIA, Novi, MI, Aug. 11-13, 2020.

1. INTRODUCTION

Armor performance requirements normally call out a specified probability of penetration and confidence level at a given velocity, for example, 90% confidence that the probability of penetration is less than 10% at 3,000 feet/second (a so-called V_{10} specification). The test procedure used to validate this performance is called out in ITOP 2-2-713 [1], and frequently results in a "22 of 22" test, in which 22 rounds are fired at the target at one specified velocity, and any single penetration invalidates the design. The result of this testing is one value for the confidence C of achieving a given level of performance P at given velocity V. The test provides no further information on how the design performs at other velocities, where it fails, or how the confidence varies with the levels of performance and velocity.

The rationale for the current testing regime is based on binomial statistics and has the advantage of extreme simplicity. However, it has a number of shortcomings in practice. For example, owing to the nature of test standards, it is common practice to test at a velocity *above* the specification, effectively using the test procedure to change the ballistic requirement. In addition, this process is sensitive to "black swan" events, where an anomaly in a single shot (for example, a round with out-of-spec properties) can invalidate an entire design.

In contrast, it is possible to use somewhat more sophisticated test procedures to ameliorate the shortcomings of the existing test regime, while at the same time acquiring substantially more information on the performance of the target with the same or fewer ballistic tests. This requires, first and foremost, a more rigorous statistical analysis of the ballistic results in order to enable hypothesis tests on the observed performance. These procedures utilize what are sometimes called "V_{50}" tests, which are simply multiple shots conducted over a range of velocities which result in both partial and complete penetrations of the target. There are a wide variety of methods to choose from in V_{50} testing, including Modified Langlie [2], Three Phase Optimal Design (3POD), Neyer Method, and Robbins-Monro-Joseph (RMJ) [3-5]. But the particular method is not as important as the statistical analysis of the results, which makes it possible to acquire an estimate of the confidence level C as a function of the probability of penetration P and the velocity V over the entire range of test conditions and beyond.

The basic concepts behind the methods described in this paper are presented in detail in two papers by Joseph Collins of the Army Research Laboratory dealing with what he calls Quantal Response, ARL-TR-6022 and ARL-TR-7088 [6-7]. The author apparently believes that we have actually *regressed* in our ability to conduct meaningful ballistic testing. The following is a quote from ARL-TR-7088, *Quantal Response: Estimation and Inference* [7] (emphasis added):

> The ballistic limit test is an important part of characterizing armor performance or establishing the lethality of projectiles. At a minimum, the analysis of such a test provides a single number which is an estimate of the ballistic limit itself. By 1950, however, it was already established that statistically meaningful decision support requires additional methodology that manifests as confidence intervals and hypothesis tests on the ballistic limit. *Unfortunately, as time passed, this knowledge was ignored.*

This paper will address the methods described by Collins and apply them with examples to V_{50} ballistic data. In addition, we will show the results of a Monte Carlo simulation of roughly 1,000 ballistic records generated from an underlying Gaussian distribution to illustrate the utility and accuracy of these methods applied to ballistic testing.

2. V_{50} TESTING AND LIKELIHOOD ESTIMATES (LE)

The main concept behind the statistical analysis of V_{50} test data involves calculation of likelihood estimates. Consider the ballistic data shown in Table 1 below, derived from a 25-shot sequence. The columns show the test velocity together with the penetration result. In addition, for each shot a probability of penetration is shown. This is based on a Gaussian (probit) distribution in which the probability is calculated from the cumulative probability density function with a given mean μ and standard deviation σ.

Table 1: V_{50} Ballistic Test Data

Test Data			
Velocity	**Result**	**P(Pen)**	**P(Result)**
3,303	Complete	93.9%	93.9%
3,064	Partial	0.0%	100.0%
3,138	Partial	0.0%	100.0%
3,232	Partial	1.6%	98.4%
3,375	Complete	100.0%	100.0%
3,271	Partial	45.5%	54.5%
3,374	Complete	100.0%	100.0%
3,236	Partial	2.7%	97.3%
3,328	Complete	99.8%	99.8%
3,318	Complete	99.0%	99.0%
3,275	Complete	53.7%	53.7%
3,218	Partial	0.2%	99.8%
3,295	Complete	87.1%	87.1%
3,241	Partial	4.7%	95.3%
3,245	Partial	7.2%	92.8%
3,249	Partial	10.5%	89.5%
3,287	Partial	76.3%	23.7%
3,276	Partial	55.8%	44.2%
3,294	Complete	86.0%	86.0%
3,250	Complete	11.4%	11.4%
3,290	Complete	80.8%	80.8%
3,296	Complete	88.2%	88.2%
3,235	Partial	2.4%	97.6%
3,262	Partial	28.1%	71.9%
3,289	Complete	79.4%	79.4%

$$P_{PEN}(V_{TEST}) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{V_{TEST}} e^{\frac{(V-\mu)^2}{2\sigma^2}} dV \qquad (1)$$

The probability of obtaining the result shown is equal to either P_{PEN} or $1–P_{PEN}$, depending on whether the shot was a complete or partial penetration. The likelihood of obtaining the observed test results (assuming each event is independent) is simply the product of the probability of each individual result. This value $L(\mu,\sigma)$ is called the likelihood estimate, and is a function of the values of μ and σ and the test data. For a probit/Gaussian distribution, the value of μ corresponds to the V_{50} velocity at which we would expect half of the rounds to penetrate the target. The value of σ is identified with the standard deviation. Figure 1 below shows a contour plot of the likelihood function over the (V_{50}, σ) plane.

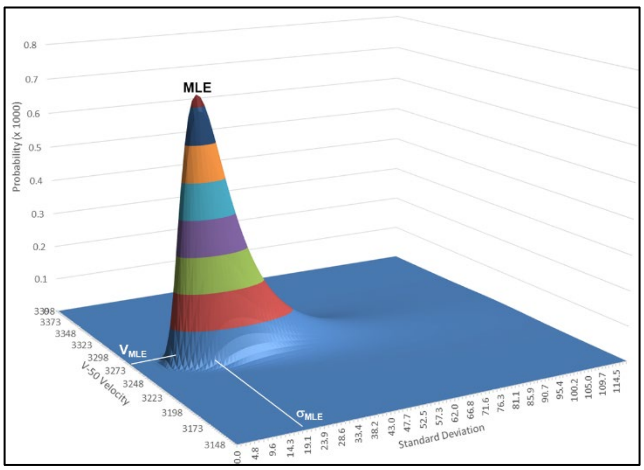

Figure 1: Likelihood as a function of V_{50} velocity and σ.

The peak of this distribution occurs at the values of V_{50} and σ which maximize the likelihood function. These are called the "maximum likelihood estimates", or MLE values. In this case, these are $V_{MLE} = 3{,}273$ and $\sigma_{MLE} = 19.3$. With these values, it is possible to calculate the probability of penetration at any velocity. So, for example, we can find the velocity at which the probability of penetration equals 10%, the so-called V_{10} velocity, which is $\Phi^{-1}(10\%, 3273, 19.3)$, where Φ^{-1} is the inverse of the cumulative distribution function with the given probability, mean, and standard deviation, which gives $V_{10} = 3{,}248$.

This analysis allows us to plot the data on a curve showing the test results together with the MLE probability of penetration, as shown in Figure 2.

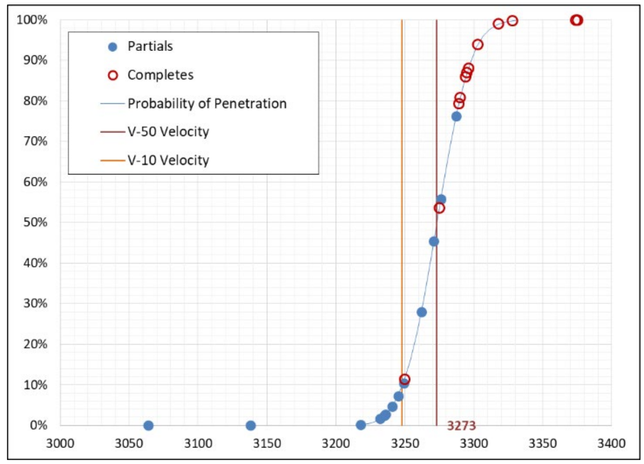

Figure 2: Test data plotted with MLE P_{PEN} estimate.

While the MLE curve provides a relation between P_{PEN} and velocity, it does not provide what we need for validation – a confidence estimate. How certain are we that the test data supports the assertion that the actual V_{10} value is greater than or equal to that calculated from the MLE distribution? That is, how confident are we that we meet the requirement? This is where hypothesis testing comes into play.

3. HYPOTHESIS TESTING

As mentioned earlier, ballistic requirements often specify a minimum level of performance (probability of penetration) at a given velocity, with a minimum confidence level – for example, 90% confidence of less than 10% P_{PEN} at 3,220 fps. We are concerned that we don't meet this requirement, so we can use this as our null hypothesis:

$$H_0: P_{PEN}(3{,}220 \text{ fps}) > 10\% \qquad (2)$$

And our alternative hypothesis is that we actually do meet spec, so:

$$H_1: P_{PEN}(3{,}220 \text{ fps}) \leq 10\% \qquad (3)$$

The question now is what test statistic do we use to decide between these hypotheses? Collins presents the generalized likelihood ratio, *GLR*, which is defined as the ratio of the likelihood of a given condition divided by the maximum likelihood over all possible conditions. In this case, the values of V_{50} and σ in the probit model of the distribution determine the conditions. That is, there are some regions of the (V_{50}, σ) plane over which the probability of penetration is more than 10%, corresponding to the null hypothesis where we fail to meet the ballistic specification. The probability that the actual values of V_{50} and σ fall in this region depends on the maximum value of the likelihood ratio L_{MAX} in this region, compared to the maximum likelihood value, *MLE*, over the entire (V_{50}, σ) plane, $GLR = L_{MAX}/MLE$.

Figure 3 below shows a plan view of the contour plot shown earlier, indicating the regions where we meet or fail to meet the spec, as well as the *MLE* value.

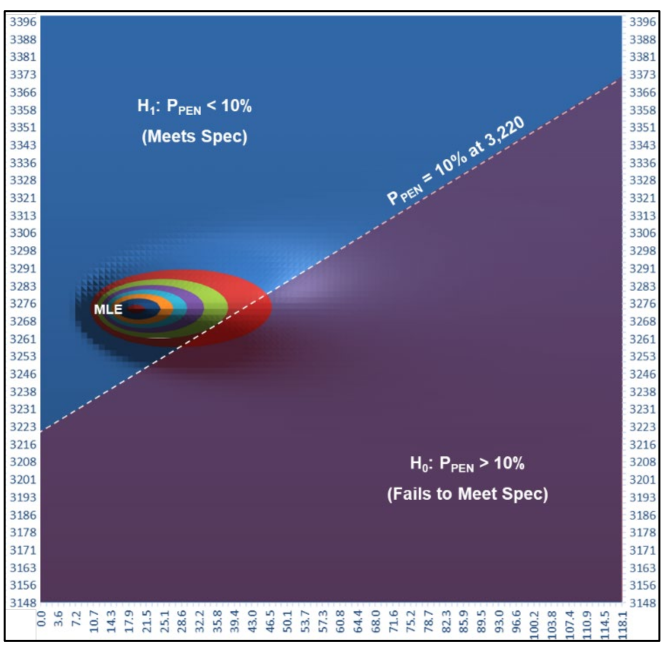

Figure 3: Contour plot of likelihood function over the (V_{50},σ) plane, showing the regions where we meet or fail to meet spec.

The dividing line in Figure 3 cleanly separates the two hypothesis regions of the plane. The equation for the line is found from inverting equation (1) above, so that we have:

$$V_{50} = V_{SPEC} - \Phi^{-1}(P_{PEN}) \cdot \sigma \qquad (4)$$

Where $\Phi^{-1}(x)$ is the inverse of the cumulative normal distribution function. Below this line, H_0 is valid. Figure 4 below shows the two likelihood values, L_{MAX} and *MLE*, used to calculate the generalized likely ratio in a contour plot similar to that shown in Figure 1 with the region of the null hypothesis cut away.

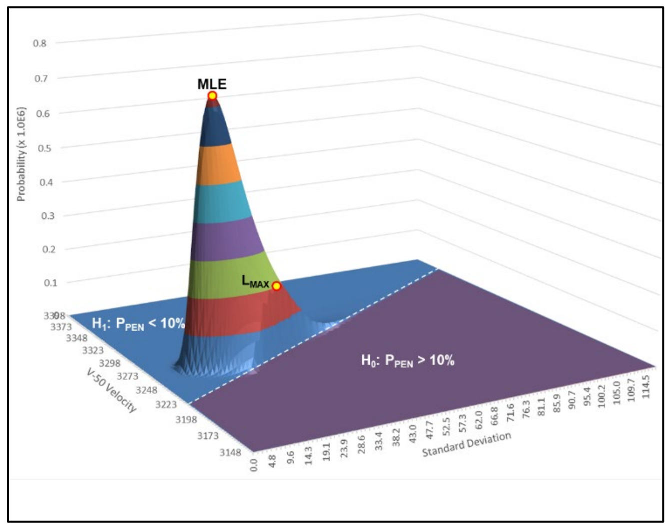

Figure 4: Evaluation of L_{MAX} using the likelihood function and the ballistic performance specification.

According to Collins [6-7], the appropriate hypothesis test statistic is the deviance Δ, which is defined as:

$$\Delta = -2 \cdot \ln(\mathrm{GLR}) = -2 \cdot \ln\left(\frac{\mathrm{L_{MAX}}}{\mathrm{MLE}}\right) \qquad (5)$$

This test statistic should be distributed according to a two-tailed χ^2 distribution. The likelihood model has two degrees of freedom because of its two parameters, V_{50} and σ. The cumulative χ^2 with two degrees of freedom happens to be a particularly simple function:

$$\chi^2_{CDF}(\Delta) = 1 - e^{-\frac{\Delta}{2}} = 1 - \alpha \qquad (6)$$

(Note that this is the area to the left of Δ, while most χ^2 charts show the area to the right, denoted by the symbol α). Also, we are only concerned with a one-tail distribution rather than a two-tail – we are only worried if the performance is too low, not too high – so the α-value is actually one-half that of the regular χ^2 [7]. So, the probability that we can reject the null hypothesis – the confidence that the probability of penetration actually meets spec – is simply:

$$Confidence = 1 - \frac{GLR}{2} = 1 - \frac{L_{MAX}}{2 \cdot MLE} \qquad (7)$$

This is fairly easy to calculate. For a given set of ballistic data, all that is required is to calculate the MLE value and then find the maximum likelihood along the line corresponding to the spec, equation (4). The confidence of meeting spec is then given directly by equation (7).

4. EXAMPLE CALCULATIONS

Consider the data shown in Table 1 above. The MLE mean and standard deviation are 3,273 and 19.3. Suppose the spec calls for a 10% probability of penetration at 3,220 – how confident are we that the data supports this assertion?

If we use the MLE values for mean and standard deviation, we find that the velocity at 10% P_{PEN} is 3,248, which is better than 3,220. But that's not the question – we need to know the confidence. If we look at equation (4), we need to find the largest likelihood value consistent with the condition that:

$$V_{50} = 3{,}220 - \Phi^{-1}(10\%) \cdot \sigma \qquad (8)$$

The value of $\Phi^{-1}(10\%)$ is –1.28155. The ratio of the maximum likelihood along this line on the (V_{50}, σ) plane to the MLE value is 0.26, occurring at (3266, 35.95) so the confidence that we meet spec is 1 – 0.13, or 87%. Note that this calculation provides the confidence at a single point, the one called out in the requirement. But the method can generate confidence at any combination of velocity and P_{PEN}. Figure 5 shows the confidence as a function of P_{PEN} for a range of striking velocities. The 87% confidence at 10% P_{PEN} with a velocity of 3,220 is just one of the points on these curves.

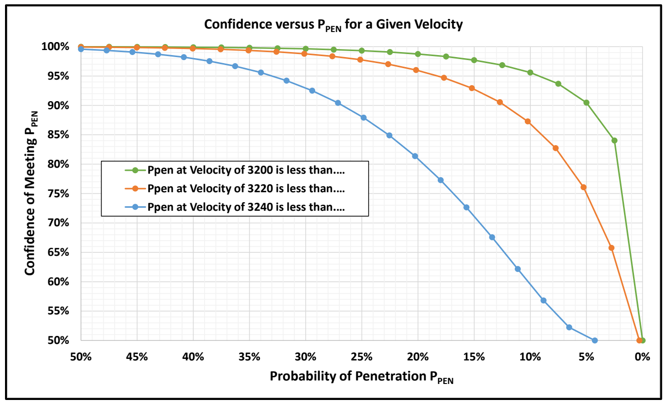

Figure 5: Confidence versus P_{PEN} for multiple striking velocities.

Similarly, we can create a chart showing the confidence as a function of striking velocity for various levels of performance. This is shown in Figure 6 below.

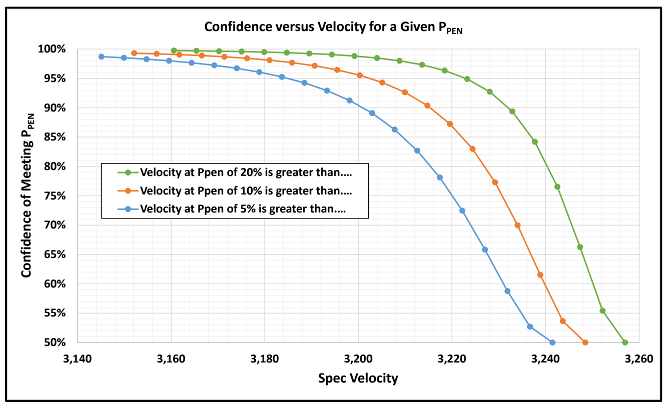

Figure 6: Confidence versus specified striking velocity for multiple values of P_{PEN}.

Again, for comparison, the legacy "22 for 22" test regime provides only a pass/fail criterion for a single point on the charts shown above.

5. MODEL VERIFICATION

The statistical analysis described here provides much more information than a simple pass/fail evaluation of a target. However, at this point it is entirely a mathematical exercise. In order to have confidence in the analysis, we need to see it applied to a wide range of data and its conclusions evaluated against the known performance of a given armor design. However, there are no "known" armor performance guidelines against which we can benchmark the model. There is no standard armor target and round combination for which the penetration versus velocity curve has been mapped with the necessary degree of precision. There is, however, a way to conduct this verification – through simulated ballistic testing against a "perfect" penetration-velocity curve modeled as a cumulative Gaussian distribution with a specified μ and σ. The goal here is to conduct simulated V_{50} tests, analyze the data using the statistical methods described here, and check to see whether they predict confidence levels in accordance with the known underlying distribution.

For this purpose, we created a simulated threat/target combination with a probability of penetration described by μ=3,000 and σ=30. We then conducted approximately 1,000 simulated V_{50} tests using a Modified Langlie method together with a random number generator to determine the result of each simulated "shot". We performed 22-shot simulated V_{50} tests, in which the first shot of each test was selected at random from a uniform distribution centered on the actual V_{50} value. We then culled the test records that had no zone of mixed results, since a ZMR is very helpful in calculating the MLE value. This left 995 valid test records simulating V_{50} tests conducted against a "perfect" target.

The question we want to address is the extent to which the analysis predicts confidence levels that are in accord with the data from the underlying distribution. For this purpose, consider the chart displayed in Figure 7. This shows the confidence versus striking velocity curve for P_{PEN}=10% for one of the 995 test records, derived from a simulated V_{50} test with 22 shots. Each point on the curve shows how confident we are that the actual V_{10} velocity is higher than the value on the abscissa. On this same chart, we show the underlying cumulative Gaussian distribution from which the data was generated, together with an indicator showing the actual V_{10} velocity, which is 2,962.

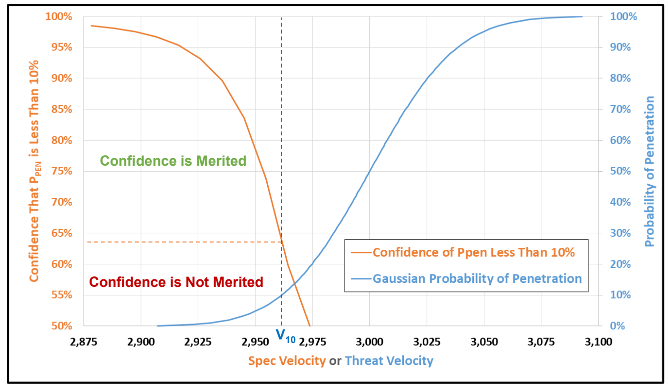

Figure 7: Confidence versus velocity for P_{PEN}=10% for one simulated test record, and underlying distribution.

The figure shows that for this test record, our confidence in the V_{10} limit is merited when the confidence level is over about 64%. For confidence levels below this, our confidence is not merited. For example, we are about 55% confident that the V_{10} is greater than 2,970, but in fact the V_{10} is less than that. On the other hand, we are 90% confident that $V_{10} \geq 2{,}935$ which is correct. If our statistical analysis is valid, then when we check to see whether we are correct about the V_{10} value at any given confidence level, we should be right about as often as our confidence. That is, if we do the analysis of 995 test records, then the actual V_{10} value should be higher than the V_{10} value at the 90% confidence level about 90% of the time.

This is a simple analysis to undertake, and can be conducted using desktop software (Microsoft Excel 2013™). The results are given in Figure 8, which shows the frequency with which the actual V_{10} velocity (within 1 fps tolerance) was greater than the lower limit of the V_{10} velocity estimated from the simulated data records at a given confidence level. The results indicate very close agreement, with a root mean square error of about 0.5% between the confidence levels and the frequency with which the calculated V_{10} levels are correct. The chart extends down to 70% confidence, spanning the range over which ballistic specifications are often written.

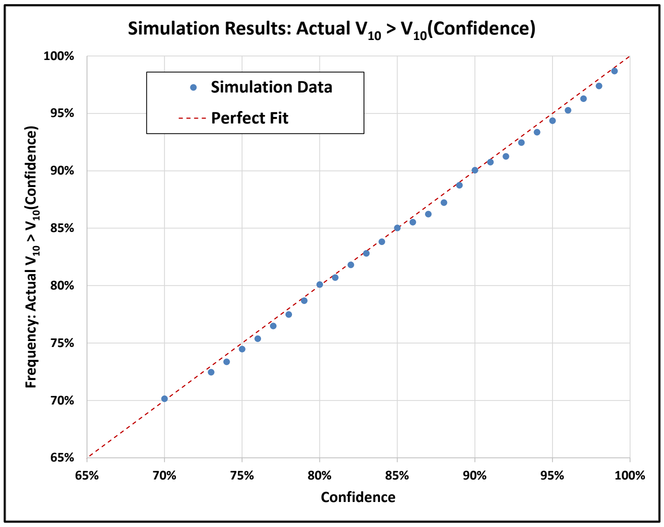

Figure 8: Results from 22-shot simulations showing the confidence of the V_{10} estimate versus the frequency with which it is correct.

Although the agreement between the confidence model and the simulated Monte Carlo ballistic data appears very good, there is a problem – the model used here is not the same as that described in the literature. In Collins' work [6-7], the number of degrees of freedom of the cumulative chi-square distribution used in hypothesis testing is one, rather than two as used in the analysis shown above. As a consequence, the equation recommended in Collins' work for the confidence is actually:

$$Confidence = 1 - \frac{[1 - \chi^2(\Delta,1)]}{2} \qquad (8)$$

The difference between a chi-square with one degree of freedom as opposed to two is significant. It is a simple matter to re-compute the analysis using one degree of freedom, and the results are shown in Figure 9 below which compares the results with one degree of freedom versus two. As can be seen, the one degree of freedom model does not match the expected confidence of the underlying distribution. Worse, the analysis is not conservative. That is, the frequency with which our analysis matches the predictions of the underlying distribution is significantly less than the expected confidence.

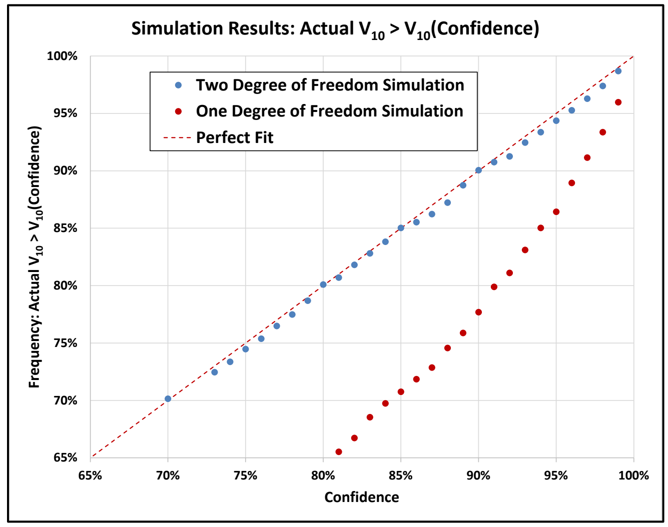

Figure 9. Comparison of simulation results using one degree of freedom versus two in the chi-square hypothesis test.

The mismatch between predicted and actual confidence using the standard one degree of freedom analysis has been noted by others. In particular, Roediger [9] says the following:

> Critical decisions are often based upon confidence bound estimates. Better guidance on their use for decision makers would be welcomed. Some experts hold the view that confidence bounds in the sensitivity testing setting can't be taken too seriously. On the other hand, using confidence bounds as in choosing a D-Optimal phase II test sequence is a useful application of them.

6. LIMIT VELOCITY ESTIMATES

The Modified Langlie approach used in producing simulated ballistic records is designed to focus on the V_{50} velocity, rather than a required probability of penetration such as a V_{10}. As a consequence, it sometimes results in ballistic records with a narrow zone of mixed results. In examining the 995 test records, it was found that many of them had a very narrow ZMR, with only about 4 fps between the slowest complete and the fastest partial. This tends to result in a higher estimated V_{10} bound at a given confidence level, owing to an abnormally low value of σ. Alternate test procedures, such as RMJ, Neyer, or 3POD, purposely test further away from the V_{50} velocity in order to focus on higher performance goals such as a V_{10}. This may increase the separation between fast partials and slow completes, reducing the likelihood of test records with a narrow ZMR, but may also skew the distribution of chosen test velocities.

While Figure 8 shows the potential accuracy of the two degree of freedom method, it remains to be seen how closely the calculated limit velocity corresponds to the actual limit. For this purpose, we looked at the difference between the calculated and actual V_{10} velocity for each test record, and created a histogram showing the frequency of these values for all 995 tests. The results are shown in Figure 9, which indicates that, for example, about 48% of the simulations provided a 90% confidence V_{10} value within 50 fps below the actual limit. The blue bars in the figure are those for which the velocity was below the actual V_{10} (the confidence was merited), while the red bars indicate those for which the velocity was higher than the actual V_{10} (the confidence was not merited).

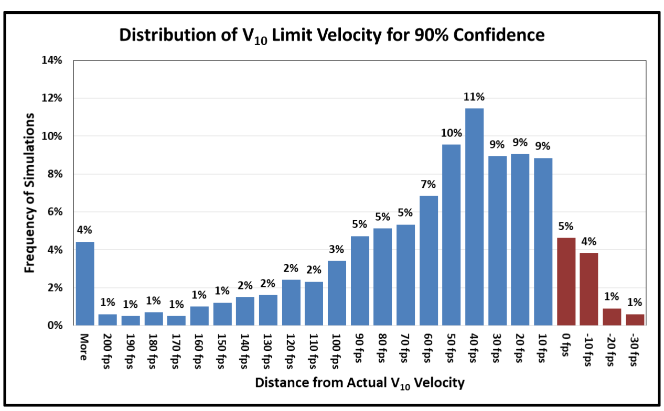

Figure 10: Distribution of error in V_{10} limit velocity for 90% confidence level.

The same analysis can be conducted at lower confidence levels, which results in pushing the limit velocity closer to the actual V_{10} value. Figure 10 shows a similar histogram for the case of 70% confidence.

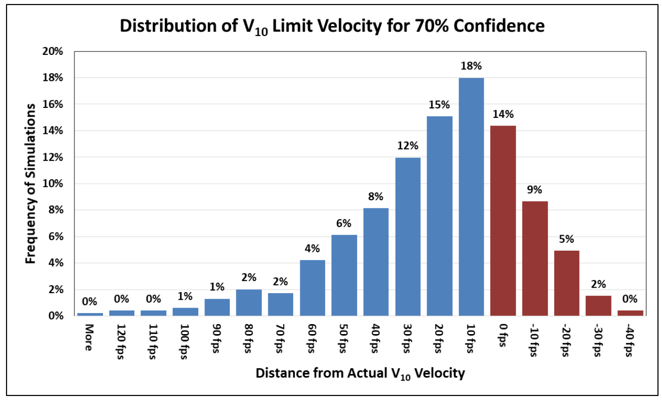

Figure 11: Distribution of error in V_{10} limit velocity for 70% confidence level.

7. CONCLUSION

The ballistic test and analysis procedures described here rely on traditional statistical methods for hypothesis testing using straightforward calculations that can be handled on any desktop computer. The Monte Carlo simulations show the accuracy of the two degree of freedom confidence level calculations using simulated ballistic test data, providing results that agree closely with the underlying distribution. The information provided by this method includes a map of the confidence of meeting required ballistic performance as a function of projectile velocity and probability of penetration over the entire range of interest. In contrast, the legacy "22 of 22" test provides only a pass/fail evaluation at a single point.

In addition, the test procedure described here is much less sensitive to a number of factors that can improperly invalidate an armor design. For example, shots fired from a test weapon naturally vary in velocity, so some tolerance about the desired spec velocity is allowed during validation. Owing to test rules (fast partials are valid shots, slow partials are not), most testing occurs well above the spec velocity, sometimes at the upper limit of the tolerance band. This effectively changes the ballistic specification, and can result in invalidating a design that actually meets specification. Because a single penetration results in failure, armor concepts are frequently over-designed, resulting in heavier packages that place more strain on other vehicle systems.

In contrast, the validation method described here tests over the entire velocity range, with multiple partial and complete penetrations. A single complete does not automatically disqualify an armor design. In addition, there is no velocity tolerance to maintain – there are no "too fast" or "too slow" invalid shots – they all count. This fact also helps dilute the influence of outlier shots. For example, a single out-of-spec test projectile may penetrate a target at a lower velocity. In the legacy test regime, this disqualifies the armor design. In the alternative validation testing described here, it results in just another complete to be analyzed with the rest of the data.

As noted above, the choice of the test algorithm can affect the likelihood of having a reasonable zone of mixed results. The Robbins-Monro-Joseph (RMJ), three-phase optimal design (3POD), and Neyer test procedures are all designed to focus on a particular ballistic requirement, for example a V_{10}, as opposed to a V_{50}. It would be interesting to see an analysis of a similar Monte Carlo test performed with these different procedures, compared to the Modified Langlie method used here. In addition, it would be interesting to see how the confidence versus V_{10} velocity curves are affected by the choice of procedure and by the number of shots in each test record. One would expect that the more shots, the tighter the tolerance. That is, the difference between the actual V_{10} velocity and the 90% confidence V_{10} velocity should get smaller as the number of shots increases. This could also be investigated through Monte Carlo simulations.

The approach described here was validated through simulations using an underlying Gaussian (probit) model of ballistic performance. Other distributions are also commonly used in analyzing ballistic data, such as the logistic (logit) or complementary log-log models. A National Institute of Standards and Technology study [8] examined the differences between these distributions when evaluating ballistic data taken from body armor testing and found no significant preference for one over the other – all three seemed to work well. It would be interesting to see whether a difference in the underlying distribution would make any significant difference in the results of Monte Carlo simulations of confidence estimates.

Most importantly, it would be good to further examine the difference between hypothesis testing conducted with a single degree of freedom chi-square hypothesis analysis, versus a two degree of freedom analysis. In theory, the one degree of freedom analysis should be the proper approach, as it corresponds to the difference between the number of degrees of freedom in the denominator and numerator of the generalized likelihood ratio. However, simulations show that using this analysis is less accurate than a two degree of freedom model, and not conservative. Further investigation of this issue could be undertaken with additional simulations.

8. REFERENCES

[1] U.S. Army Aberdeen Test Center, "ITOP 2-2-713, Ballistic Testing of Armor," DTIC ADB289615, page 10, 2002.

[2] U.S. Army Aberdeen Proving Ground, "TOP 2-2-710, Ballistic Tests of Armor Material," DTIC ADA137873, 1984.

[3] C.F.J. Wu and Y. Tian, "Three-phase optimal design of sensitivity experiments," J. Statist. Planning Inference, vol. 149, pp. 1-15, June 2014.

[4] B.T. Neyer, "A D-optimality-based sensitivity test," Technometrics, vol. 36, no. 1, pp. 61-70, February 1994.

[5] V.R. Joseph, "Efficient Robbins-Monro procedure for binary data," Biometrika, vol. 91, no. 2, pp. 461-470, 2004.

[6] Joseph C. Collins, “Quantal Response: Practical Sensitivity Testing,” ARL-TR-6022, U.S. Army Research Laboratory, 2012.

[7] Joseph C. Collins, “Quantal Response: Estimation and Inference,” ARL-TR-7088, U.S. Army Research Laboratory, 2014.

[8] Diane Mauchant et al., “Analysis of three different regression models to estimate the ballistic performance of new and environmentally conditioned body armor,” NISTIR 7760, National Institute of Standards and Technology, 2011.

[9] Paul A. Roediger, “GONOGO: An R Implementation of Test Methods to Perform, Analyze, and Simulate Sensitivity Experiments,” AREIS-CR-18007, U.S. Army Armament Research, Development, and Engineering Center, 2018.