Shielded from Oversight

The Disastrous US Approach to Strategic Missile Defense

http://www.ucsusa.org/shieldedfromoversight

Appendix 8: Confidence Levels and Probability

Union of Concerned Scientists

Why Getting a Good Estimate of the Kill Probability Requires Many Tests

For a system such as a missile defense in which the outcome of a test can be evaluated as either a success (destruction of the target) or a failure (the target gets through), the probability of each outcome can be described by a binomial distribution.

The binomial distribution may be familiar; it describes situations such as a coin toss, which also has two outcomes: heads and tails. The probabilities of heads (p) and tails (q) sums to 1.0. (If the coin is unweighted, heads and tails are equally likely, so the probability p of heads is 0.5 as is that of tails.)

For a binomial distribution, the probability P of getting exactly k successes (heads, for example) in n trials (coin flips) is

$$P(k) = \frac{n!}{k!\,(n-k)!} p^k (1-p)^{n-k}$$

As discussed further below, it's important to note that the binomial distribution holds only if the probability of success, p, is the same for every trial and if the trials are statistically independent (the outcome of one trial does not affect the outcome of the next.)

To use the binomial distribution to estimate how effective a missile defense system is, one can describe the probability of success (the "kill probability") as p and of failure (the "miss probability") as $1 - p$ for a single shot. Using the binomial distribution, in a set of n incoming missiles, each targeted by an interceptor with a kill probability p, the probability P that k warheads penetrate the defense is:

$$P(k) = \frac{n!}{k!\,(n-k)!} (1-p)^k (p)^{n-k}$$

And the probability that no (zero) warheads penetrate the defense is

$$P(0) = p^n$$

Therefore, the probability that at least one warhead will get through the defense in an attack of n missiles is

$$1 - P(0) = 1 - p^n$$

For strategic missile defense, the expectation is that the system must be highly effective. However, even with a kill probability p of 95 percent, the probability that at least one warhead will get through in an attack of five missiles (using one shot on each incoming missile) is perhaps surprisingly high, almost one in four:

$$1 - P(0) = 1 - 0.95^5 = 0.23$$

In reality, the GMD system is complex and requires a number of major systems not only working well on their own but also working well together, and under conditions expected to be challenging and variable. Thus, a single shot kill probability of 0.95 would be very difficult to achieve.

Additionally, this simple approximation, using a single value for kill probability and assuming failures are uncorrelated is not completely realistic, as the kill probability will depend not only on the specific hardware (the interceptor variant, for example), but on the conditions under which the shot is taken. In addition, there may be sources of common error, such as a component that always fails in each interceptor or the presence of enemy countermeasures that confuse the defense. Also, the presence of one interceptor may affect the success of the other interceptors, for example, by confusing the sensors with additional signals from the kill vehicle's steering thrusters or debris from the interceptor.

Confidence Intervals: How Well Do You Know the Value of *p*?

It is critical not just to have an estimate for the value of p but also to understand how good your estimate is, for reasons elaborated in the next section. The kill probability p must be estimated from tests (and simulations of tests), much in the same way one would have to do a number of flips to determine whether a coin was weighted to one side or if both

sides were equally likely to turn up. The task of determining the quality of the estimate is a bit more difficult than one might expect. For example, if you flip an unweighted coin 100 times, and then repeat that process many times, you would find that heads would come up exactly 50 times in 100 flips only about 8% of the time.

How many tests do you need to do? The number depends both on the underlying *p* and how well you want to know it. As a guide to intuition, we again assume the simplest scenario: the tests are performed under the same conditions, the underlying *p* is the same in each trial, and the test results are uncorrelated to each other. (This is actually not the case for the existing suite of Ground-based Midcourse Defense (GMD) tests. For example, the tests did not use the same sensor information to determine the intercept point, and many different variants and subvariants of the interceptor have been used in the 17 intercept tests.)

In this simple case, if a set of *X* successes are observed in a set of *n* tests, the estimated value of the kill probability *p* is $\hat{p} = X/n$. The more tests are performed, the closer $\hat{p}$ will be to the true (but unknown) value of *p*.

Without knowing the true value of *p*, how do you know how accurate $\hat{p}$ actually is? A statistical measure of this is the confidence interval, which is the range of values that will include the true value of *p* a given percentage of the times that $\hat{p}$ is estimated from a set of *n* tests. For example, the 68-percent confidence interval will contain the true value of *p* 68 percent of the times that $\hat{p}$ is estimated, and the 95-percent confidence interval will contain the true value of *p* 95 percent of the times that $\hat{p}$ is estimated using that method. The confidence interval will be narrower (will span a smaller range of values) the larger the number of tests, *n*, that have been conducted. The size of the confidence interval also depends on the underlying probability distribution.

While the binomial distribution is simple in form, calculating confidence levels for $\hat{p}$ for this distribution is more complex. The simplest approach is the Wald approximation: approximating the distribution of errors as the same distribution as errors for a normal, or Gaussian, distribution. The Wald approximation has the advantage of providing an easily calculated form for the confidence interval. If $z_{\alpha/2}$ is the $100(1 - \alpha/2)$th percentile of the standard normal distribution, then the area under the standard normal curve (centered on zero) to the left of the value $z_{\alpha/2}$ will be $(1 - \alpha/2)$. Using $z_{\alpha/2}$ (which can be read off a standard normal distribution table) and $\hat{p} = X/n$, the kill probability estimate, the $100(1- \alpha)$th percent confidence interval centered around $\hat{p}$ can be found:

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{1}{n}\hat{p}(1-\hat{p})}$$

To calculate a two-sided 95-percent confidence interval (i.e., the true value is between the upper and lower boundary values at the 95-percent confidence level), $z_{\alpha/2} = 1.96$.

To calculate a one-sided confidence interval, (i.e., the true value of *p* is greater than some value at a given confidence level), that value, i.e., the lower bound on the confidence interval is:

$$\hat{p} - z_{\alpha}\sqrt{\frac{1}{n}\hat{p}(1-\hat{p})}$$

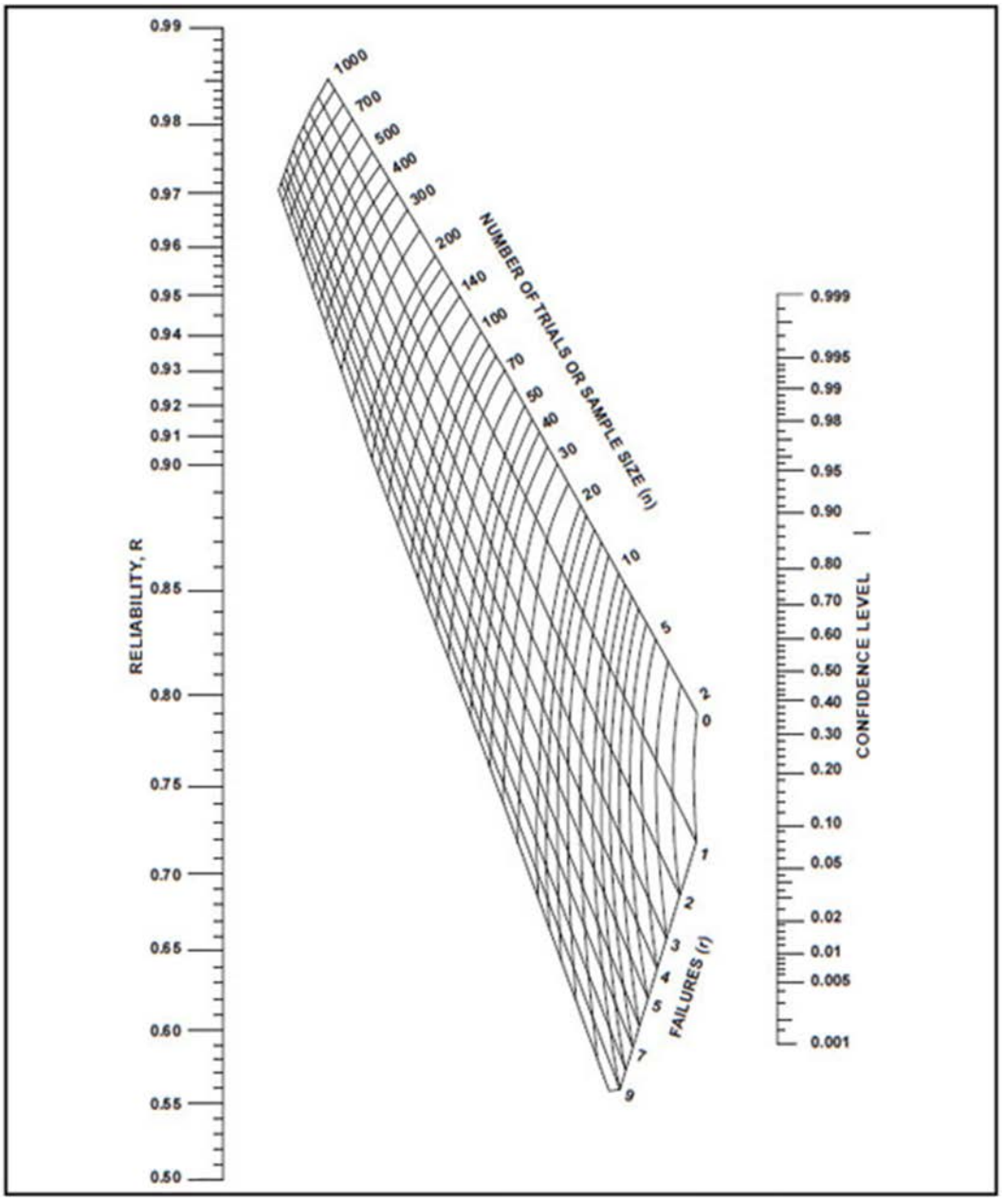
For a 95% confidence, interval, $z_{\alpha} = 1.645$.

However, the Wald approximation becomes less accurate at values of *p* that are close to 0 or 1, and has some complex behavior at other values.[1] Other measures of the confidence interval have been developed, such as the Agresti-Coull approach, which provides better estimates than the "normal" distribution approximation, but is still simple in form. It is sometimes called "add two successes and two failures" because that is how it differs from the Wald approximation:

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{1}{\tilde{n}}\tilde{p}(1-\tilde{p})}$$

[1] For an overview of interval estimation for the binomial proportion and evaluation of different approaches, see Brown, L. T. Cai, and A. DasGupta. 2001. Interval estimation for a binomial proportion. *Statistical Science* 16(2):101–117. Online at *http://www-stat.wharton.upenn.edu/~lbrown/Papers/2001a%20Interval%20estimation%20for%20a%20binomial%20proportion%20(with%20T.%20T.%20Cai%20and%20A.%20DasGupta).pdf*. Note: All URLs in footnotes to this appendix were accessed May 24–25, 2016.

FIGURE 1. Nomogram of the cumulative binomial distribution. The two axes on the graph are the number of tests (n) and the number of failures (r). Any straight line that goes through a point on this graph will connect an estimate of the reliability (R), analogous to $\hat{p}$ in this discussion to confidence level at which it is known. Source: Defense Acquisition University.



where the "new" number of trials $\tilde{n} = n + 4$ and the "new" estimate of the kill probability is $\tilde{p} = (X + 2)/(n + 4)$.

To take an example, if a system had 17 identical tests with 8 successes (the record for GMD intercept tests given in Shielded From Oversight: Table 1: Ground-based Midcourse Defense Intercept Tests [2]), the Wald approximation gives an estimate of $\hat{p}$ to be 0.47 ± 0.24 at 95 percent confidence, or $0.23 < \hat{p} < 0.71$. The reader should treat these kill probability estimates with caution, of course, since the tests were not performed under the same conditions, and so were not testing the same thing. (See Appendix 7: Testing.) The one-sided 95-percent confidence interval for such a test record is $\hat{p} > 0.27$. The Agresti-Coull 95-percent confidence interval estimate for the same test record would be 0.47 ± 0.21, or that the kill probability p is greater than 0.29 at 95% confidence. On the one hand, these kill probability estimates quickly tell the observer that the GMD system is not very reliable; on the other hand, they also say that the reliability is not characterized very well.

Another way to present the relationship between the test record, estimated reliability, and confidence level is in the form of a nomograph or nomogram, a graphical calculating diagram. Figure 1 is the nomogram for the cumulative binomial distribution.[3] The intersection of the lines representing the number of tests and number of failures produces a point; any straight line through this point will connect the value of the reliability with the confidence level at which it can be expressed. For example, one can say with about 30 percent confidence that a system with nine failures in 17 tests has an estimated reliability of at least 0.50, or has a chance of hitting a target (under those exact conditions) at least 50 percent of the time.

To turn the question around the other way, then: how many tests (n) are needed to estimate p, the kill probability under a specific set of circumstances, to a useful amount of precision? While it would take 30 tests without a failure to be 95- percent confident that the system was at least 90-percent reliable, on the other hand, one can tell the system is not very reliable with a relatively few number of tests if they fail half the time.

One may want to know not just the lower bound to the confidence interval. Rather, the two-sided confidence interval may be useful, for reasons such as those discussed in the next section. The precision D, which is half of the width of the confidence interval, can be written using the Wald approximation:

$$D = z_{\alpha/2}\sqrt{\frac{1}{n}\hat{p}(1-\hat{p})}$$

It is inverted easily, giving the estimate:

$$n = \hat{p}(1-\hat{p})\frac{z_{\alpha/2}^2}{D^2}$$

for a level of precision D. So for a two-sided 95- percent confidence interval that is 0.20 wide (i.e., the precision D = 0.10) for an actual underlying kill probability of p = 0.5 (i.e. 50 percent), 96 tests would need to be completed. (This estimate is fairly consistent with more rigorous methods of calculating n.)[4] That is, to establish the kill probability's 95% confidence interval is between 0.40 and 0.60, around 100 tests under the same conditions are necessary, as shown in Table 1 below.

Note that the precision measure D is half the width of the confidence interval, not a percentage of $\hat{p}$. So one cannot, for example, use this estimate to find the value of $\hat{p}$ with a precision of D = 0.20 if $\hat{p} = 0.10$ or 0.90. If the actual kill probability were much lower or much higher, for example, p = 0.10 or 0.90, the number of tests needed would be fewer, but still substantial, i.e., 35.

[2] The GMD intercept test record from Shielded from Oversight's Table 1 differs from the Missile Defense Agency's assessment by one success. Since the interceptor in FTG-02 struck the target but did not destroy it, we do not consider this to be a successful intercept test.

[3] This nomograph is sourced from the Defense Acquisition University's Program Manager's Toolkit, an online set of resources derived from classes for Department of Defense staff. Online at *https://acc.dau.mil/CommunityBrowser.aspx?id=294528.*

[4] Compare with Table 2 in Krishnamoorthy, K., and J. Peng. 2007. Some properties of the exact and score methods for binomial proportion and sample size calculation. *Communications in Statistics—Simulation and Computation* 36: 1171—1186. Online at *www.ucs.louisiana.edu/~kxk4695/com_stat_bin_07.pdf*.

TABLE 1. The number *n* of tests required to determine the kill probability *p* with a precision *D* at 95-percent confidence, using the Wald approximation. If the underlying probability is, for example, *p* = 0.90, 138 tests would be required to determine to 95-percent confidence that *p* is between 0.85 and 0.95.

p	*D*	*n*
0.10	0.05	138
0.10	0.10	35
0.50	0.05	384
0.50	0.10	96
0.90	0.05	138
0.90	0.10	35

time-consuming to perform, the number of actual intercept tests conducted will never be great enough on their own to provide high confidence in the estimated kill probability under the different circumstances the system is expected to perform. Instead, the Missile Defense Agency relies on computer simulations, which are anchored by data generated in flight testing and ground testing. This approach has significant limits (see Chapter 4 of Shielded from Oversight.)

The Institute for Defense Analyses, tasked with carrying out an assessment of the GMD system, pointed out that the Pentagon does not know the system's reliability very well at all and probably will not in the future either. It notes that due to the limited number of flight and ground tests there is "a significant degree of uncertainty in the estimates of current and projected GBI reliability" and that the tests planned for the future "are likely to be insufficient by themselves to reduce significantly this uncertainty."[5]

Why a Good Estimate of Kill Probability is So Important

For a strategic missile defense system, which is meant to defend against nuclear weapons, a robust knowledge of the system's capabilities is necessary for making informed decisions about how much it can be relied on in a crisis as well as decisions about what kind of resources to invest in it.

The kill probability is an important component of understanding how many warheads would be expected to penetrate the defense in an attack. A useful model for the single-shot probability of kill (SSPK) was developed in Wilkening[6], and the following is adapted from it.

The probability that one or more warheads will leak through a defense is $1 - P(0)$, where $P(0) = (K_W)^W$ is the probability that no warheads get through in an attack of W warheads, and K_W is the kill probability against one warhead. K_W depends on the probability, P_{track}, that the system can track and identify the target reliably and that it does not have common mode failures, as well as on the SSPK, analogous to *p* above. The number of shots taken on a given warhead is *n*, thus the kill probability against one warhead is:

$$K_W = P_{track}(1 - (1 - \text{SSPK})^n)$$

And

$$P(0) = [P_{track}(1 - (1 - \text{SSPK})^n)]^W$$

Assuming for these purposes that $P_{track} = 1.0$, meaning the system has a perfect ability to track and identify targets accurately among other objects, then we can investigate how the probability that the defense leaks at least one warhead depends on the SSPK in different scenarios. (The probability that no warheads would get through decreases by a factor of $P_{track}{}^W$ in this construction.)

For example, in the situation in which all 44 of the planned GMD interceptors were to be used against a raid size of 11 apparent warheads (with four-on-one targeting), a defense with perfect P_{track} and an SSPK of 0.50 would let at least one warhead through the defense with a probability of about 50 percent.

[5] Institute for Defense Analyses. 2012. *IDA's responses to questions on the "Independent review and assessment of the Ground-Based Midcourse Defense system. Paper P-4802."* Portions unclassified. April 11

[6] Wilkening, D.A. 1999. A Simple Model for Calculating Ballistic Missile Defense Effectiveness. *Science and Global Security* 8(2): 183-215. Online at http://scienceandglobalsecurity.org/archive/2000/01/a_simple_model_for_calculating.html.

If the SSPK were lower than 0.50 —for example, if SSPK is equal to the lower limit of the 95-percent confidence interval for the set of 17 tests described in the first section of the appendix, with SSPK = 0.23—then the defense would perform poorly, even using four on one targeting. Indeed, it would, with 99-percent probability, let through at least one warhead in such an attack. Using an optimistic estimate for the SSPK, the upper limit of the confidence interval, 0.71, the defense would let at least one warhead through with an 7.5 percent probability, a very different outcome. Clearly it is crucial to know as much as possible about the interceptor's reliability.

At present, the actual SSPK is undoubtedly quite low, judging from the test record. While using more than one interceptor against a target can make up for poor performance, this strategy is not effective until the SSPK is fairly high. The number of interceptors that can be reasonably targeted on a given warhead is not arbitrarily high. Wilkening suggests a cap of four-on-one targeting is likely and that adding more interceptors may provide diminishing returns in a crowded and confusing field.
In the scenario of an attack of five warheads, it is clear that four-on-one targeting cannot make up much ground for ineffective interceptors; for an SSPK of 0.25, four-on-one targeting will let through at least one warhead 85 percent of the time, as shown in Table 2 below. For interceptors with an SSPK of 0.50, at least one warhead will get through a fourth of the time in such an attack. It is only when the SSPK is relatively high that using multiple interceptors on a target can provide an effective defense; four-on-one targeting with a SSPK of 0.90 will defeat all incoming warheads 95 percent of the time. Also note that one-on-one targeting of the highly effective system with an SSPK of 0.90 would let through a warhead 40 percent of the time in an attack of five warheads.

Note that a shoot-look-shoot scheme doesn't affect the probability that warheads survive the defense over simply targeting the warhead with multiple interceptors, as long as the defense has enough interceptors to fire the desired number at each target. What it can do is reduce the number of interceptors used to achieve a given effect, and therefore conserve a limited inventory for use against future attacks. Shoot-look-shoot will not improve the system's effectiveness, but it can improve its efficiency.

TABLE 2. The probability that at least one warhead survives the defense in an attack of five warheads, given the targeting scheme (the number of interceptors targeted on each warhead), and the single shot kill probability SSKP.

Targeting scheme	Single shot kill probability	Probability at least one warhead survives
1-on-1	0.10	99.99%
2-on-1	0.10	99.98%
4-on-1	0.10	99.5%
1-on-1	0.25	99.90%
2-on-1	0.25	98%
4-on-1	0.25	85%
1-on-1	0.50	97%
2-on-1	0.50	76%
3-on-1	0.50	49%
4-on-1	0.50	28%
1-on-1	0.90	41%
2-on-1	0.90	5%
4-on-1	0.90	0.05%