A Significant Improvement for Anti-malware Tests

Richard Ford and Marco Carvalho
Harris Institute for Assured Information
Florida Institute of Technology
Melbourne, FL 32901
(rford, mcarvalho)@fit.edu

Abstract—**Despite ongoing improvements in the quality of antimalware tests, the way in which test results are handled often shows a low level of sophistication. In this paper, we introduce the simple concept of confidence intervals and statistical significance to these tests, and show that many of the "best practice" approaches common in other fields are lacking in the security-software testing industry. Further, we argue that the lack of these techniques harms the industry as a whole, and provide a road map for broader adoption of well-known statistical techniques for estimating the confidence interval on measurements.**

I. Introduction

There was once a time where anti-malware tests consisted of a tester, often with unstated qualifications, obtaining a set of malicious files, blindly running scans against them, and publishing what was missed and what was detected. Usenet was littered with enthusiastic amateurs who would make measurements, write reports, and draw conclusions. While these days are (thankfully!) long gone now, the steady maturation of anti-malware testing techniques needs to continue, evolving with the needs of the test consumer community and the ever-changing threat landscape itself.

In this paper, we focus on just one aspect of anti-malware testing: the uncertainty surrounding the right way to interpret test scores. Using basic probability and error analysis, we show how the confidence we can place in a particular result varies as a function of the number of samples tested against. Using these statistical methods, we turn our attention to current generation anti-malware tests: to what extent do existing tests meaningfully discriminate between the efficacy of different products?

II. A Brief History of Tests

Anti-malware Testing has a long history — for as long as there have been anti-virus products, there have been those who test those products. However, unlike traditional products such as a Word Processor, security products have critical functionality that requires specialist knowledge to test. For example, an anti-malware product's most important functionality is to protect the user from malicious code… but determining how well a product does this is fraught with challenges.

At the simplest level (and as reflected by early anti-malware tests) one takes a large collection of malware and measures what is detected. Naively, the product that "detects more samples" is the best product. Such hubris, however, is not backed up by science, as even this seemingly simple test is anything but straightforward.

First, how the samples are selected is incredibly important. With the number of malicious samples skyrocketing, it is not practical to test on *all* possible samples; instead, a subset of samples are selected for testing. These samples should be representative of the overall threat level if the test result claims to have predictive power with respect to the product's real-world efficacy.

Second, it is not enough to merely measure detection; the tester is in fact interested in the Type I and Type II error rates of the product — or, as we call these numbers, False Positives and False Negatives. Indeed, in the early days of anti-malware testing, Dr. Alan Solomon would famously argue how a simple batch file that makred all samples as infected would score perfectly on tests that did not handle these issues correctly [1].

As if this were not enough, there is another significant complication, as vendors began making use of the infection vector and the cloud; for the former, many vendors now look at *where* a sample comes from. For the latter part, most anti-malware products now have a significant cloud component where intelligence about a particular file can be used to drive, or at least weight, detection decisions. With inter-vendor back-end file sharing, a sample shown to Product A may be already known to Product B even if it is a new sample, if A shares with B in real time. Such tests must therefore carefully weigh infection vector and test timing.

As tests become more complex, they tend to become more expensive to carry out, and so there is a tension between the size of the test corpus and the cost of the test. For simple "scan everything on a hard drive" tests, it is relatively cheap to scan thousands – or even tens of thousands – of samples. However, for complex tests such as malware removal or APT detection, testers sometimes test on very small numbers of samples to keep costs down. Sometimes, the number of samples can be as low as ten (see, for example, [2]). In such cases, the challenge of statistical significance rears its ugly head, and it is to this problem that we now turn.

III. How many samples matter?

The question of how many samples one really needs to test in order to know, with a certain degree of certainty, that product A is likely to detect more samples than product B sounds trivial, and indeed, mathematically, it is. However, for non-statisticians (a group that makes up a surprisingly large portion of the anti-malware industry), the results can seem counterproductive.

When we think about errors in test results, we need to think about two different types of error: systematic error, and random error. Systematic error is the kind of error that stems from incorrect measurement techniques. For example, let us consider a physical example. If we are measuring the resistance of a circuit and our measuring equipment is incorrectly calibrated, *all* our measurements may be incorrect. Similarly, if our resistance meter suffers from drift as a function of time (perhaps as it warms up), repeated measurements of a physical quantity will show a corresponding change as a function of time. In the anti-malware world, a trivial case of systematic measure might be a tester who obtains all its test samples from a single vendor. This will skew the test results toward the sample provider, and is an example of systematic error brought about by poor test procedures. This idea of the impact of sample selection being important to testing is not new. AMTSO has published a set of guidelines for sample selection, and this goes a long way to removing systematic error, at least that which stems from sample selection. However, the presence of random error is not in any way removed.

When we consider random error, we must consider errors that are unpredictable, and that are scattered around a true value. Using our resistance example from above again, a random error might be the result of random fluctuations in the effectiveness of our measurement apparatus. A property of random error is that the more times a measurement is repeated, the more accurate the average (mean) result is.

With the difference between random and systematic error firmly in our minds, we now consider the effect of random error on the accuracy of anti-malware software tests. For this discussion, we will assume that the tests are carried out correctly and fairly; that is, if sample X is reported as detected by Product A, this is a true and correct result; a similar case holds for samples that are missed. Our thesis — and we are on very solid scientific foundations here — is that the security testing industry has almost universally ignored the existence and impact of random error on test design, execution, and interpretation. This has led to misleading results and unsupported conclusions on the part of testers or test consumers.

A. *One Bag, Many Balls...*

The easiest way to illustrate the challenges raised by random error are with an example drawn from basic probability and statistics. Once the idea is grasped with day to day objects with which we are familiar with, it is much easier to see how it applies to the more abstract world of malware.

For our *Gedanken* we imagine a small bag filled with marbles. The marbles may be red or black, but we do not know the makeup of the bag initially; that is, we know it contains some number of red and some number of black marbles, but we do not know the ratio of marble colors in the bag.

Let us now suppose we draw just one marble from the bag and look at it, and let us assume it is red. Nobody would assume that we have learned much of anything about the contents of the bag, except that there is at least one red marble initially in there. Now, let us draw another nine marbles... and each of these is *also* red. We now have good evidence we can use to make inferences about the rest of the bag. Assuming that our drawing was truly random, we might conclude that the bag contains *no* black balls, but a quick pause for thought will allow us to back away from such a conclusion. We might think that the chances are strong that the number of red balls dramatically outweighs the number of black balls, but it is too much of a step to say that there are *no* black balls in there. Similarly, it is possible that we drew red balls purely by chance. If there are 99 black balls for every red ball and if the bag is large enough that the number of balls is essentially infinite, the chances of picking a single red ball are:

$$p_{red} = \frac{1}{100} \tag{1}$$

As independent probabilities are multiplied, we can then calculate the probability that despite there being predominantly black balls in our bad, we draw all red marbles as:

$$p_n = \left(\frac{1}{100}\right)^n \tag{2}$$

where n represents the number of balls we draw, and p_n is the probability each ball is red. A tiny probability, we agree, but also not impossible.

Looking at these simple equations, we can see that the chances of drawing all red balls goes down the more balls we draw. That is, the more times we measure the system, the lower the probability of our combined measurements being an outlier. Put simply: the more times we examine the contents of the bag by randomly selecting a ball, the better *confidence* we have in our conclusion.

The question then becomes given a set of observations, what does this set tell us about the Universe as a whole? In our previous example, what did our experiment tell us about the likely makeup of the entire bag's content? Similarly, we can ask, what does a test on a subset of malware tell us about a product's likely performance on the much larger universe of malware in general? How many samples do we need to test to have tested *enough*? How big a difference in detection lets us say Product A is better than Product B with some level of confidence?

B. *Calculating Confidence*

In the experimental sciences the concept of some kind of error bar, where the bar tells the reader something about the uncertainty in the measured quantity. However, there are two common types of error bar: those that show standard error (often in terms of σ, the standard deviation), and those that show a confidence interval. Note that these two similar tools tell us different things: one tells us about the standard deviation, and the other relates to the likelihood that the confidence interval has captured the true underlying value. Confidence intervals are commonly drawn such that there is a 95% chance the confidence interval has captured the "real" value.

When we are comparing experiment results, if two results have confidence intervals that significantly overlap, we cannot say with anything approaching certainty, that the two results are different. Conversely, if two measurements' confidence intervals do not overlap at all, then we have a statistical basis to assert that the two results are different.

TABLE I. CLOPPER–PEARSON INTERVALS ON THE AV-TEST DETECTION RESULTS

Product	Score	Upper CI (.95)	Lower CI (.95)
Microsoft	92 (67%)	74.4	58.1
Others	138 (100%)	100	97.2
Sophos	132 (96%)	98.3	90.6

It turns out the mathematics required to calculate an actual confidence interval is quite complex, but fortunately we do not have to do much of the underlying research: the problem is well known. In this paper, we elect to use a *binomial distribution* for our model of the number of successes/failures we are likely to see. This is actually an approximation, as the binomial distribution assumes that each yes/no test (or in the case of malware, pass/fail) is statistically independent. This would be true if we selected a sample to test randomly, and that subsequent tests could also select that sample; in fact, in anti-malware testing, testers typically choose a set of samples; there is zero chance a sample could be selected twice. Despite this, provided that the universe of possible samples, N, is much larger than the number of samples tested, n, the binomial distribution is a good choice.[1]

With this as a background, a common method of calculating confidence interval is known as the Clopper-Pearson method [3]. This method is complex to derive, but with the advent of modern computing, the approach is attractive, because it tends to overestimate the size of the confidence interval, allowing us to err on the side of conservatism, rather than over-interpret insignificant differences.

IV. WELL KNOWN TESTS, REDUX

In this section we take what we have learned, and we calculate confidence intervals based on sample size. For the purposes of these calculations we assumed that the number of samples tested was small compared to the large corpus of known malware (of whatever type we are testing).

Let us start with an easy example.

AV-Test has a section on their website for protection [4]. Looking at Windows 8.1 tests, the following products and scores are listed in Table I. As can be seen, the while most products tested scored 100%, Sophos scored 96% in April. This *sounds* pretty bad, but how significant is it? Taken alone, that equtes to 132.5 out of 138 samples (we are not sure what this fractional score means, so we chose to round the number down to be most pessimistic). Calculating the confidence interval at the 95% level, we show a range of 98.3–90.6. This interval overlaps with that of the products that scored perfectly (100–97.2) and so we cannot safely assert that Sophos' results are meaningfully worse based on this datapoint alone.

Another good example of where statistical analysis would have helped is in a test by Anti-Virus Comparative [5]. Here, removal tests were carried out using just ten samples. What makes the analysis difficult, however, is that the samples were not randomly selected, but carefully selected by hand, drawing from a set of samples that was prevalent in the wild. This manual selection is good, in that it makes the test more representative, but very difficult from an analysis point of view, because, as the samples are no longer independent or random, we cannot use a binomial distribution. However, it is possible to use the variability of scoring to calculate the standard deviation of the mean score for each product, which is of limited assistance in assessing the variability of results. Nevertheless, this test highlights how simple changes in tests can *dramatically* complicate statistical analysis.

TABLE II. ANALYSIS OF DTL'S SCORES — WITH CONFIDENCE INTERVALS

Product	Defended	Neutralized	Compromised	Protected
Symantec	98 (96–100%)	0 (0–2%)	2 (0–4%)	98 (93–99%)
Kaspersky	94 (91–99%)	3 (0–8%)	3 (0–8%)	97 (92–99%)
McAfee	91 (87–97%)	5 (1–11%)	4 (0–10%)	96 (90–99%)
Trend	79 (72–87%)	15 (8–23%)	6 (0–14%)	94 (87–98%)
Microsoft	35 (25–46%)	24 (14–35%)	41(31–52%)	59 (49–69%)

Finally, we examine a test by Dennis Technology Labs (DTL), titled "Enterprise Anti-Virus Protection" [6]. This test is particularly interesting because it highlights many of the subtle challenges involved analyzing results.

In this test, DTL elected to classify samples into three categories. A sample could be either Defended, Neutralized, or Compromised. As we now have three possible outcomes from a test, we can no longer use a Binomial Distribution, but must instead use a *Multinomal Distribution*. We can calculate how much our experiment tells us in such a case using the methods outlined by [7]. DTL's raw data is shown in Table II. 95% Confidence Intervals are provided next to the raw numbers; as is so often the case, these ranges seem very large in comparison to the differences that we have previously considered significant.

A. *Statistics for Certifications?*

Interestingly, the need for error bars disappears when we are measuring an absolute quantity, and when we are sure of our numbers. As such, those testers who engage in certification schemes do not run afoul of statistical methods, because they are typically reporting on an absolute quantity. For example, consider a certification that is based upon detecting 100% of the samples in the upper part of the WildList. For such a test, a product either passes or fails; it is absolutely reasonable to state that a product that misses one of these samples has not passed the test, and therefore should not be certified.

While this may feel somewhat counterintuitive, in fact, a momentary reflection makes it clear that this works because the entire set of samples has been tested, and therefore there is no sample selection issue related to random sampling. Had the certifier chosen, for example, 50% of the possible samples, there is room for sampling error, and the difference between 100% detection *of the 50% tested* and a single miss is negligible.

Significance and uncertainty still have a role in examining product certification. For example, a product that has 1 miss over 12 months in terms of certification is not significantly better than one which has had none; it could just be chance and random sampling. Perhaps counter-intuitively, this small sample size (12 months worth of certifications) requires a large

[1] In this paper, we have deliberately chosen to provide a simplified description of the problem; we are well aware of the challenges imposed by the less-than random sample selection process, for example, that is common in tests.

difference in results in order to be confident that the two products are different in a meaningful way.

V. Conclusion

To a mathematician, the ideas presented in this paper are in no way new or innovative, and so to one skilled in the arts, it is legitimate to wonder why we even took the time to write it. After all, one has to question what new we have to offer here? Our argument — and it is borne out by the current industry practices — is that despite the techniques described here being well known, they are either not known or deliberately not applied in the area of anti-malware testing. As such, its publication and dissemination is both meaningful and important: by increasing the rigor of anti-malware testing practices, we help exert a positive pressure on the industry toward tests that in turn exert a pressure on vendors to change their products. If this pressure is in the right direction, it will help improve protection globally; if not, it may be neutral or, at worst, detrimental to the overall protection provided to users.

The practice of highlighting small differences in product performance as significant, despite their being no mathematical basis for such statements has to stop. Understanding the accuracy of our statements, and how this can be extrapolated to tell readers something about the world at large, is an obvious next step as anti-malware testers move from *ad hoc* practices to the practice of science.

To be very clear, we are not faulting or singling out particular testers in this paper. We fully understand the tension between the number of samples tested and the cost of a test, and that testers face strong pressures to keep costs down in order to cover more ground. Conversely, we argue that such decisions must be based on science; the idea of presenting confidence intervals is unquestionably best practice in other disciplines. It is time that the industry as a whole recognizes that the significance of measurements is critical in interpreting what can be small differences in performance. To ignore this encourages gaming of tests, flawed science, and preys on those who are unaccustomed to looking at raw data.

While many of the techniques we have described her can be put in place with almost zero cost, their implications are far reaching with respect to test design. At best, we will see testers either have to admit that their tests only weakly discriminate between products, or extend their tests to a larger number of samples. While increased test set sizes are fairly trivial for some kinds of tests, for other tests, such an increase increases the cost of testing almost linearly.

Acknowledgment

The authors would like to thank Simon Edwards of Dennis Technology Labs for access to test data and helpful conversations that led us to be sure that this paper would be beneficial in the industry at large.

References

[1] A. Solomon, “Private Communication,” 1992.

[2] A. Comparative, “Malware Removal Test,” http://www.av-comparatives.org/removal-tests/, 2013, [Online; accessed 8-July-2014].

[3] C. J. Clopper and E. S. Pearson, “The Use of Confidence or Fiducial Limits Illustrated in the Case of the Binomial,” *Biometrika*, vol. 26, no. 4, pp. 404–413, 1934. [Online]. Available: http://dx.doi.org/10.2307/2331986

[4] AV-Test, “Homepage,” http://www.av-test.org/en/home/?avtest[type]=1, 2014, [Online; accessed 8-July-2014].

[5] AntiVirus Comparative, “Retrospective/Proactive test — Heuristic and behavioural protection against new/unknown malicious software,” http://av-comparative.org, 2014, [Online; accessed 8-July-2014].

[6] Dennis Technology Labs, “Enterprise Anti-Virus Protection — Jan – March,” 2014.

[7] C. P. Sison and J. Glaz, “Simultaneous confidence intervals and sample size determination for multinomial proportions,” *Journal of the American Statistical Association*, vol. 90, no. 429, pp. 366–369, 1995.