On the Statistical Properties of the F-measure

Tsong Yueh Chen
tchen@it.swin.edu.au

Fei-Ching Kuo
dkuo@it.swin.edu.au

Robert Merkel *
rm Merkel@it.swin.edu.au

School of Information Technology
Swinburne University of Technology
John St, Hawthorn 3122, Australia

Abstract

The F-measure - the number of distinct test cases to detect the first program failure - is an effectiveness measure for debug testing strategies. We show that for random testing with replacement, the F-measure will be distributed according to the geometric distribution. A simulation study examines the distribution of two adaptive random testing methods, to study how closely their sampling distributions approximate the geometric distribution, revealing that in the worst case scenario, the sampling distribution for adaptive random testing is very similar to random testing. Our results have provided an answer to a conjecture that adaptive random testing is always a more effective alternative to random testing, with reference to the F-measure. We consider the implications of our findings for previous studies conducted in the area, and make recommendations to future studies.

Keywords: F-measure, testing effectiveness metric, quality measurement, adaptive random testing, random testing, software testing.

1. Introduction

Debug testing describes the testing of a computer program (the program under test) for the purpose of detecting as many program faults as can be found, on the basis that detection and rectification of these faults will achieve the greatest possible program reliability improvement with the available resources. The key criterion for assessing debug testing strategies is therefore effectiveness - how many program faults can be detected with a given testing strategy. However, such a description is too vague to serve as a measurable statistic to compare different debug testing methods, so several statistics have been devised for just this purpose. Most make the simplifying but reasonable assumption that the cost of testing is a constant factor of the number of tests performed, and therefore compare detected failures with tests conducted.

The most common metrics used to compare testing effectiveness have been the P-measure and E-measure. The P-measure is defined as the probability that at least one program failure is detected with a specified sequence of tests. The E-measure is defined as the expected number of failures detected by a sequence of tests. They have been used to compare partition testing to random testing analytically [5] [6] [9] [10]. Other relationships between these metrics have also been determined, one of the more notable being that when they are small (≪ 1), the P-measure and E-measure closely approximate each other.

More recently, Chen et al. [4] have proposed a new effectiveness measure, the F-measure, using it to compare their black-box testing method, Adaptive Random Testing (ART) to random testing. Adaptive Random Testing was a black-box test case selection strategy based around the assumption that tests spread evenly through the input domain were more likely to detect failure than tests clumped together. The method they used to spread the test cases out was titled Fixed Sized Candidate Set-ART (FSCS-ART). FSCS-ART involves generating a fixed number of candidates and choosing the candidate most distant to already executed tests.

The F-measure is defined as the number of test cases required to detect the first failure. They argue that this measure is more informative, and natural, than the earlier measures. This metric has been adopted for subsequent studies of related work, notably Chan et al.[2], which examined Restricted Random Testing, another ART method using the principle of exclusion to achieve an even spread of test cases, and Chen et al. [3].

If the input domain of the program under test is of size $D$, and of those, $d$ inputs fail to produce the correct output, the failure rate $\theta = \frac{d}{D}$. For random testing as described previously, if $n$ tests are conducted the E-measure is $n\theta$, and...
the P-measure is \(1 - (1 - \theta)^n\). For the E-measure, it is trivial to use the Central Limit Theorem to show that the sampling distribution will be normal. For the F-measure, Chan et al. [2] have stated that the expected F-measure for random testing is \(\frac{1}{\theta}\).

In this paper, we examine the sampling distribution of the F-measure both from an analytic and experimental perspective. Section 2 examines analytically the statistical properties of the F-measure for random testing, showing that the sampling distribution of the F-measure for random testing follows a well-known distribution, the geometric distribution. In this context random testing describes a strategy where test cases are randomly selected from across the entire input domain, with a uniform probability of selection.

In section 3 we conduct a simulation study to examine whether the same sampling distributions are present when testing using FSCS-ART and RRT. Finally, in section 4 we discuss the implications of our results with regards to previous related studies using the F-measure, and recommendations for future comparison studies.

2. Sample distribution of the F-measure in random testing

If we test a program with failure rate \(\theta\) using a randomly selected test case, the probability of detecting a failure is precisely \(\theta\). Given this, if we test a program with a set of randomly selected test cases, the probability that the first test case in this set will detect a failure is also \(\theta\). In other word

\[
P(F = 1) = \theta
\]

(1)

If failure is detected with the first test case, for the purposes of calculating the F-measure the test sequence is complete. Therefore, the chances of a second test case being selected is \((1 - \theta)\). If that second test case, selected with replacement (that is, the previously executed test case is not excluded from the pool of possible selections) is executed, the chances of it detecting a failure is still \(\theta\). Therefore,

\[
P(F = 2) = (1 - \theta) \theta
\]

(2)

In general, the probability of the \(n\)th trial detecting failure is the probability of preceding trials not detecting failure, multiplied by the probability of this specific trial detecting failure:

\[
P(F = n) = (1 - \theta)^{n-1} \theta
\]

(3)

This probability distribution turns out to be an instance of the geometric distribution ([8], p. 176). Mosteller et al. ([8] p. 189), describe a proof that the expected number of trials before first failure (denoted by \(F\) in our specific case), is \(\frac{1}{\theta}\), and also ([8] p. 219), outlines a proof that the variance of \(F\) is \(\frac{1 - \theta}{\theta^2}\). If \(\theta\) is small, this simplifies to approximately \(\frac{1}{\theta}\) and the standard deviation is approximately \(\frac{1}{\theta}\).

The median F-measure can be approximated using the solution to the following summation:

\[
\sum_{n=1}^{x} (1 - \theta)^{n-1} \theta = \frac{1}{2}
\]

(4)

In general, there is no exact integral solution for \(x\), however, the sampling median will be the smallest whole number greater than the analytical solution to the summation:

\[
\text{median} = \left\lceil \frac{\ln \left( \frac{1}{2} \right)}{\ln (1 - \theta)} \right\rceil
\]

(5)

Table 1 shows that in most cases the ratio \(\frac{\text{median}}{\text{mean}}\) is approximately 0.7. As \(\theta\) approaches 0, the ratio \(\frac{\text{median}}{\text{mean}}\) decreases, but the effect is marginal over typical values of \(\theta\). In the limit as \(\theta \to 0\), the ratio \(\frac{\text{median}}{\text{mean}}\) approaches \(\ln 2 \approx 0.693\). Consequently, for typical values of \(\theta\), well over 50% of test runs will require fewer than the mean number of test cases to detect the first failure.

3. Simulation study of distribution in random and adaptive random testing

3.1. Introduction

While the previous section shows analytically the distribution of the F-measure for random testing, the analysis is not applicable to adaptive random testing as the probability of detecting failure is not constant, but should increase from trial to trial. However, it was not clear whether different versions of ART would have markedly different distributions, or whether other factors such as the distribution of inputs that fail, or the failure rate would affect the overall shape of the distributions. We therefore conducted a simulation study to examine the sampling distribution of the F-measure under these conditions. In our simulations, three different pa-
3.2. Method

Chan et al. [1], describe three common “patterns” of how a program fault cause distinctively-shaped region or regions in the input domain to fail when tested. These failure patterns are shown in Figure 1. The original paper discusses the types of faults that may cause these failure patterns, but for the purpose of the present paper we note that a block pattern is a single, continuous region of the input domain, a strip pattern is a block pattern elongated greatly in one direction, and a point pattern involves stand-alone points or very small regions spaced (sometimes regularly) throughout the input domain. Previous experiments have shown that ART methods show their greatest improvement in effectiveness on block failure patterns, modest improvement on strip failure patterns, and little to no improvement on point patterns.

For our simulations, block patterns were simulated by square-shaped regions of the input domain. Strip patterns were simulated by randomly choosing a point on a vertical boundary of the input domain, and a point on a horizontal boundary of the input domain, and plotting the boundary of a strip centred on them of width chosen so as to ensure the failure rate is as specified. Point patterns were created by randomly placing 10 circular regions without overlapping, whose total size was chosen so as to have the appropriate failure rate, in the input domain.

To conduct a simulation run, the failure pattern was chosen according to the particular experimental condition. Test cases were then generated using the chosen test case selection method until a “failure” was detected (the test case was within the simulated failure region), and the number of test cases required for the detection was recorded. For each experimental condition, 2000 simulation runs were performed.

The test case selection strategies compared were random testing, with replacement, and with a uniform probability of all points in the input domain to be selected; Fixed Size Candidate Set-Adaptive Random Testing (FSCS-ART) [4], using 10 candidates per test; and Restricted Random Testing [2], using an exclusion ratio of 150%. For full details on these methods, please consult the relevant papers; we provide only a brief summary here.

In essence, the FSCS-ART method is as follows: Select the first test case randomly and “execute” the test. For subsequent test cases, generate k candidates. For each candidate cj, calculate its Euclidean distance di to the nearest already executed test case. Choose the cj such that 

\[ d_j \geq d_i, \forall i \leq k \]

in other words, the candidate most distant from the previously executed test cases. In restricted random testing, a circular exclusion region is placed around each already executed test case, of size such that the total area covered (ignoring any overlap of the exclusion regions of the already executed test cases) is equal to that of the input domain multiplied by the exclusion ratio. A candidate is then randomly selected. If it lies outside any exclusion region, it is accepted for testing and the test case is executed - if not, additional candidates are generated until one lies outside all the exclusion regions. Note that the exclusion ratio can be over 100%, as the level of overlap is considerable.

Failure rates used in our trials were 0.02, 0.01, 0.002, and 0.001. With 4 failure rates, 3 different types of failure patterns, and 3 testing strategies, there were 36 experimental conditions in total.

3.3. Results

As our simulations generated a considerable amount of data, we discuss here only results for the failure rate 0.002, and due to page limit, we cannot present all results in this paper. Varying the failure rate had little qualitative effect on the shape of the distribution except the obvious scaling factors.

Figure 2 shows a histogram (smoothed by placing the data in buckets) for the three different types of failure patterns using random testing. As can be seen, the resulting distributions are identical and a good match for the geometric distribution predicted by our analysis in the previous section.

For the ART methods, Figures 6 and 7 show that the distributions for the point and strip patterns were close to those for RT. It is anticipated that if the failure rates were higher, there would be a more significant difference between the distributions of ART and RT for strip patterns. Our experimental data for \( \theta = 0.01 \) and 0.02 are consistent with this expectation. For the block pattern (Figure 5), however, the
distribution is significantly different, with a much steeper drop-off in frequency for high F-measures. Furthermore, at small F-measures, the frequency for ART is significantly higher than that for RT.

The same data, grouped by testing method, clearly shows the effect of the failure pattern on the F-measure distribution for ART methods. Figures 3 and 4 show the steep drop-off for block patterns, the close-to-geometric distribution for point patterns, and the strip patterns much closer to the point pattern in this case.

4. Discussion

In this paper, we have shown that the F-measure will have a geometric sampling distribution, and examined the properties of this distribution in the light of testing practice.

In previous studies, effectiveness has been examined purely by quoting mean F-measures. Given the distributions we have shown here, we believe that this does not provide a complete basis for comparison. With the default assumption of normality, there is an expectation that the mean, median, and mode of the quoted statistic will be the same, and that data will be clustered around the mean in the usual pattern. These assumptions are false, in this case, with the median substantially smaller than the mean.

Our simulation study showed that for block patterns, FSCS-ART and RRT produce characteristic sampling distributions that resemble a geometric distribution much more closely than the normal distribution. The distinctive feature of both distributions is a much steeper drop-off than in random testing. A much less pronounced drop-off is evidenced for strip patterns. This is consistent with previous studies that have shown a large reduction in mean F-measure using FSCS-ART and RRT on block patterns, and a much more modest reduction for strip patterns. The sampling distributions of FSCS-ART and RRT are very similar to each other. Although they are based on different intuitions to evenly spread test cases, their performance is shown here to be very similar. This characteristic drop-off indicates that for ART methods, there will be a quite large proportion of runs that detect failures very quickly, and a quite high percentage of runs below the mean.

4.1. Robustness of ART methods

One interesting observation can be made of the effectiveness of FSCS-ART and R-ART in the most unfavourable case - the point failure pattern. Previous studies have shown that in these cases, the mean F-measure is very similar to random testing. Our results show that not only are the means similar, the sampling distributions are very similar.
This provides further evidence of the robustness of the ART methods, in that, at worst, their effectiveness is similar to that of random testing.

4.2. Implications for previous related studies

Given these non-normal and somewhat divergent distributions, the reader may question whether previous reported statistical comparisons of adaptive random testing methods have provided a full and complete picture. We first consider the descriptive statistics. One of the consequences of the well-known Central Limit Theorem is that the distribution of a sample mean will tend towards normality for sufficiently large sample sizes, regardless of the distribution of the population being sampled. Therefore, confidence intervals for population means can be accurately estimated regardless of the distribution of the sampled population. The only caveat outstanding is whether the sample standard deviation is an accurate estimator of the population standard deviation; Mendenhall and Sincich ([7], p. 307) state that the approximation is generally “quite satisfactory” if the sample size is over 30. The sample sizes used in these earlier studies were over an order of magnitude greater than this, and so we believe the estimates were quite justified. The studies also quote sample standard deviations, and used them to dynamically decide whether the sample size was large enough (when the confidence interval for the mean was small enough). This also appears to be acceptable. None of the studies quoted confidence intervals for the standard deviation, however, a naive approach to doing so would have been statistically erroneous, as the standard methods for doing so assume that the sample is approximately normally distributed regardless of the sample size ([7], p. 333).

As far as inferential statistics are concerned, the main such statistics were comparisons of two sample means. Given the assumption of “large” (> 30) sample sizes (and most sample sizes were at least an order of magnitude greater than this), testing whether the difference between two sample means is statistically significant does not require any assumptions about the sampling distributions of the two populations ([7], p. 374). However, while the inferential statistics actually performed in the past are valid, other, more elaborate types of inferential statistics often have more stringent assumptions, and great care should be taken when applying them to non-normal situations.

It should also be noted that most statistically significant results quoted in past studies were highly statistically significant and the sample sizes were extremely large. The distributions of the various test case selection methods are also quite similar. All of these properties are, in general, good indicators that statistically significant results are genuine.

As well as being statistically valid, we believe that pre-
Figure 4. Histogram of block, strip and point patterns when performing restricted random testing with failure rate 0.002

Figure 5. Histogram for block patterns with failure rate 0.002
Figure 6. Histogram for strip patterns with failure rate 0.002

Figure 7. Histogram for point pattern with failure rate 0.002
viously quoted results are meaningful. While the sampling distributions of random testing and ART methods are different, they are not radically different - a lower mean F-measure from ART also corresponds to a lower median and less likelihood of a very high F-measure on a single test set.

As discussed in section 2, when failure rates (θ) are small, the standard deviation of the F-measure is about $\frac{1}{\theta}$, which is the same as the mean. Hence, our theoretical result has explained why in previous simulations, the mean and standard deviation are very close to each other.

4.3. Future Work

However, given the distinctive, and slightly unusual sampling distributions shown by FSCS-ART and RRT, it would seem that, while statistically valid, simply quoting means and standard deviations to characterise the F-measures of the methods is less than optimally illustrative. The automated nature of simulation and experimental studies in this area makes it quite straightforward to obtain very large sample sizes, and if sample sizes are large enough, confidence intervals become so small that any measurable difference is statistically significant. If this is done, the entire sampling distributions of methods can be accurately characterised and compared using graphical methods. We believe that, for the purposes of comparing effectiveness, a cumulative histogram is a particularly useful comparison tool.

Our results have concentrated on the F-measure. It is trivial to show that empirical estimates of the E-measure will have a normal sampling distribution, using the Central Limit Theorem. The sampling distribution of estimates of the P-measure has not been studied here, and is likely to be skewed, possibly following a Poisson distribution. Given the close approximation between the E-measure and P-measure when they are small, we expect the skew to reduce and approximate normality in this case. Theoretical or simulation analysis to examine this distribution further would be a worthwhile future project.

5. Acknowledgements

This project was partially supported by an Australian Research Council Discovery Grant (DP0345147).

References


