ApEn Test Parameter Selection (DRAFT)

Joshua E. Hill, InfoGard Labs
e-mail: jhill@infogard.com

December 15, 2004

Abstract: *This paper refines the meaningful range of Approximate Entropy Test block length selections for bit sequences of 1,000,000 bits. For this sequence length, the block sizes of 11-14 bits are rejected for all meaningful statistical testing, and the block sizes 9 bits and 10 bits are found to be acceptable for limited testing. The block size of 8 bits is found to be generally acceptable for use with this sequence length.*

Key words: *random number generator testing, statistical testing, randomness testing, approximate entropy test, NIST SP800-22, sts.*

1 Introduction

The NIST sts tool (a reference implementation of the statistical tests described in NIST SP800-22 [4]) implements a set of tests that verify that a stream of bits has some of the same statistical properties as a stream of bits of the same length produced by a true random number generator (RNG).

The reference document describes 16 tests, each of which provide some information about the bit string under review. In order to use these tests, the tester must partition the provided bit string into a set of bit sequences. Each bit sequence should be the same length (n bits long). Each bit sequence will undergo each of the 16 tests. Each test outputs at least one result (called a test P-value). This P-value is the test's assessment of the "randomness" of the bit sequence. (The P-value can be interpreted as the percentage of time that a true-RNG would produce a bit sequence that appeared to be "less random" than the bit sequence under test).

In order to facilitate coming to some conclusion, a value is selected as a cutoff point (generally called "α", "the critical value", or "the level of significance"). If the P-value is this value or above, the test is considered a pass. If the P-value is below this value, the test is considered a failure. The sts package has this value hard-coded into the application as $\alpha = 0.01$.

The distribution of these P-values is expected to be uniform (i.e. if a true RNG were to undergo this testing regimen, the resulting P-values would asymptotically approach a uniform distribution). As a consequence, even a true RNG would occasionally "fail" these tests (the percentage of tests that fail is expected to be α).

For each of the tests, the output test P-values are interpreted in two ways to establish if the provided bit string is evidently random; both of these methods are based on the idea

that the P-values should be uniformly distributed. The first method examines the percentage of "passing" P-values. This pass rate is expected to fall in a certain range centered about 1-α. If the results significantly deviate from this pass rate (either by failing or passing more than expected), the bit string under review is considered non-random.

The other method of interpretation involves comparing the distribution of test P-values with a uniform distribution. If the test P-value distribution looks sufficiently non-uniform, then it is considered a failure. This test operates by forming an independent statistical test which compares the observed data with a uniform distribution, resulting in a distribution P-value (called P-value$_T$), which has its own critical point set to 0.0001. If the resulting P-value$_T$ is less than this critical value, the bits under review are considered non-random.

The NIST sts tool requires parameter selection for many of the statistical tests implemented. One of these tests is the Approximate Entropy (ApEn) test, which is described in the SP800-22 document in section 2.13 and 3.13, and additionally in [1], [2], and [3]. The ApEn test requires the operator to select its block size.

The acceptable range for this block size is based on the bit sequence length. Reasonable bounds for the block size have been established through experimentation, as the theoretical estimates have been based on asymptotic assumptions, which are not necessarily consistent with the behavior of the test with "small" sequence lengths (e.g. 1,000,000 bits). This paper attempts to refine the set of reasonable ApEn block size settings for 1,000,000 bit sequence length.

2 Parameter Selection

The following test settings are consistent with SP800-22:

Configuration Item	**Setting**
Bits per bit sequence (n)	1,000,000
Number of bit sequences (sample size)	1,000+
Frequency Test within a Block block size	20,000
Non-Overlapping Template Test template length	10
Overlapping Template block length	10
Maurer's "Universal Statistical" Test block length (L), initialization steps (Q)	L=7, Q=1,280
Approximate Entropy Test block length	10
Serial Test block length	16
Linear Complexity block length	1,000

The test-suite-wide parameters are the number of bits per bit sequence (n) and the number of bit sequences. n must be selected to be consistent with the requirements all of the tests to be run. The Overlapping Template Matching Test, Linear Complexity Test, Random Excursions Test, and Random Excursions Variant Test all require n to be greater than or

equal to 10^6 in order to produce meaningful results. The Non-Overlapping Template Matching Test and the Lempel-Ziv Compression test require n to equal 10^6. (See SP800-22 sections 2.7.7, 2.8.7, 2.10.7, 2.11.7, 2.15.7, and 2.16.7) The Lempel-Ziv code in the sts tool works only if n is one of a few discrete values, of which the only meaningful setting is $n=10^6$.

The number of bit sequences (sample size) must be 1,000 or greater in order for the "Proportion of Sequences Passing a Test" result to be meaningful. (See SP800-22 section 4.2.1 and 4.3 f)

Given these two settings, settings for the rest of the tests narrow. For the Frequency Test within a Block, if $n=10^6$, the test block size should be set between 10^4 and 10^6. (See SP800-22 section 2.2.7)

The two template tests (Non-Overlapping Template Test and Overlapping Template Test) both require selection of a template length of 9 or 10 in order to produce meaningful results. (See SP800-22 section 2.7.7 and 2.8.7)

Maurer's Universal Statistical Test block length (L) and initialization steps (Q) must be consistent with the table in SP800-22 section 2.9.7. For $n=10^6$, the only acceptable values are (L=6, Q=640) and (L=7, Q=1280). Interestingly, the sts tool ignores the user-provided settings, and sets internally consistent values based on n.

Selection of the Approximate Entropy Test block length is the principle topic of this paper. SP800-22 section 2.13.7 requires the block length to be less than $\lfloor \log_2 n \rfloor - 2$, however the sts tool warns if the block size is greater than $\lfloor \log_2 n \rfloor - 5$ (which is consistent with the information in section 4.3 f). As such, SP800-22 and the sts tool indicate that the ApEn block size should be 14 or less.

The Serial Test block length is also set based on n. If $n=10^6$, the block length must be less than 17. (See SP800-22 section 2.12.7)

The Linear Complexity block length is required to be set to between 500 and 5,000, inclusive. (See SP800-22 section 2.11.7)

It should be noted that non-parameter selection problems have been reported in two of the tests in the sts test suite (the Lempel-Ziv Test and the Discrete Fourier Transform Test) [0].

3 ApEn Block Size Selection

It is the nature of the ApEn test that larger block sizes provide better information about the bit sequence being tested, as long as the block size is not set so large that the test becomes meaningless. Unfortunately, the theoretical expected bound for this parameter is misleading: The expectation is that a block size approaching $\lfloor \log_2 n \rfloor$ would be

acceptable, but NIST has established empirically for n=1,000,000 that values greater than 14 begin to disagree with the expected value (SP 800-22 section 4.3 f).

We initially chose a value of 10 bits in order to assure that the ApEn test remained meaningful while still extracting good information from the test. This selection was consistent with the block size selected by NIST for testing [5] (see footnote 6 on page 2).

On Sep 10, 2004, Jan Blonk of TNO pointed out that failures seemed fairly common for the selections of an ApEn block size 13 and 14 for n=1,000,000, after a meaningful number of trials (over 1,000 trials). We ran a series of tests internally, and concluded that the ApEn block sizes of 13 and 14 indeed gave results that did not agree with the expected statistic (in particular, the distribution of ApEn test P-values was insufficiently uniform), even when testing known good generators.

4 Testing

Testing was conducted using a modified version of NIST's published sts-1.5 test tool. The pertinent modifications were as follows:

- Replaced the non-compliant FIPS 186-2 PRNGs with a compliant implementation (note: the sts tool's ANSI X9.17 generator is also non-compliant)
- Added the ability to use a configuration file to allow for automation of these tests
- Added the ability to specify RNG seeds in the configuration file
- Added the ability to perform data analysis on previously generated results

We used a reference FIPS 186-2 PRNG [6] to produce bit sequences for analysis. The FIPS 186-2 standard describes a family of related PRNGs; the employed PRNG is completely described as a FIPS 186-2 Appendix 3.1 general purpose PRNG, using the SHA-1 based G function. This PRNG was validated as correct using the NIST CAVS testing tool.

Each tested bit sequence was 1,000,000 bits long (n=1,000,000). Every thousand rounds of testing, a P-value$_T$ was produced using all of test P-values calculated up to that point. (The first data point is the P-value$_T$ for the first thousand samples, the eleventh data point is the P-value$_T$ for the first eleven thousand samples, etc). Testing for each block size occurred until either the tests persistently failed or 1,000,000 rounds of testing were complete, whichever occurred first.

The idea underlying this style of testing is that the reference generator (the identified FIPS 186-2 PRNG) should necessarily pass eventually. As such, the test should eventually converge to a pass, but it may display fluctuations in its results before the number of tests becomes statistically significant. The statistic that was most problematic was the distribution of ApEn test P-values, so our testing concentrates on this statistic. In particular, if the distribution of the test P-values generally becomes less uniform over a statistically significant number of samples, we take this as an indication that the parameter selection is invalid (i.e. if P-value$_T$ decreases consistently, the test parameter selection is considered invalid).

This strategy resulted in the testing of 10^{12} bits (1 trillion bits) for the ApEn block sizes of 8 and 9. Testing for the block size of 10 bits was discontinued after approximately 170,000 rounds of testing (170 billion bits). Testing for the block sizes 11 and 12 each underwent 10,000 rounds of testing (10 billion bits), and block sizes 13 and 14 each underwent 1,000 rounds of testing (1 billion bits). In each case, the initial starting seed was set to be the same value (XKEY= 5e892383a8e7c9fb32c9fdcf2abd44e5b0554d14), and so the same bit sequences were tested in each case.

5 Results

As previously noted, the ApEn block sizes 13 and 14 persistently failed after less than 1,000 rounds. The block sizes 12 and 11 persistently failed within 10,000 rounds of testing. The block size 10 failed after approximately 170,000 rounds of testing. The block size 9 did not persistently fail within the testing, but its distribution P-value did trend downwards. Based upon this trend, it would appear that the ApEn block size 9 should fail in fewer than 2,000,000 rounds of testing. The ApEn block size of 8 did not show any downward trend. See Figure 1 (log scale) and Figure 2 (linear scale) for a representation of the distribution P-value$_T$ throughout the testing.

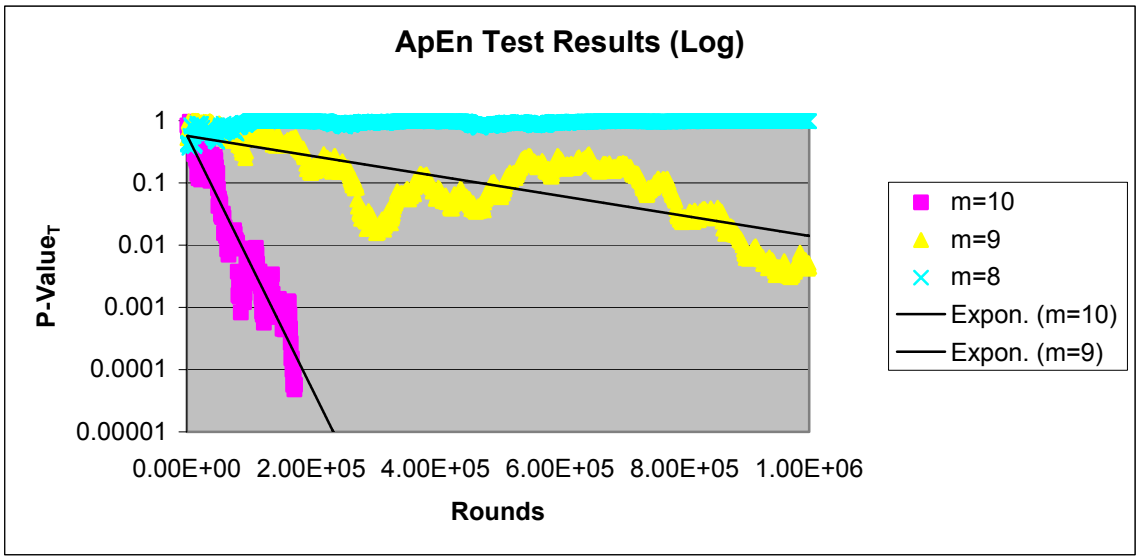

Figure 1: P-value$_T$ vs. number of rounds (log)

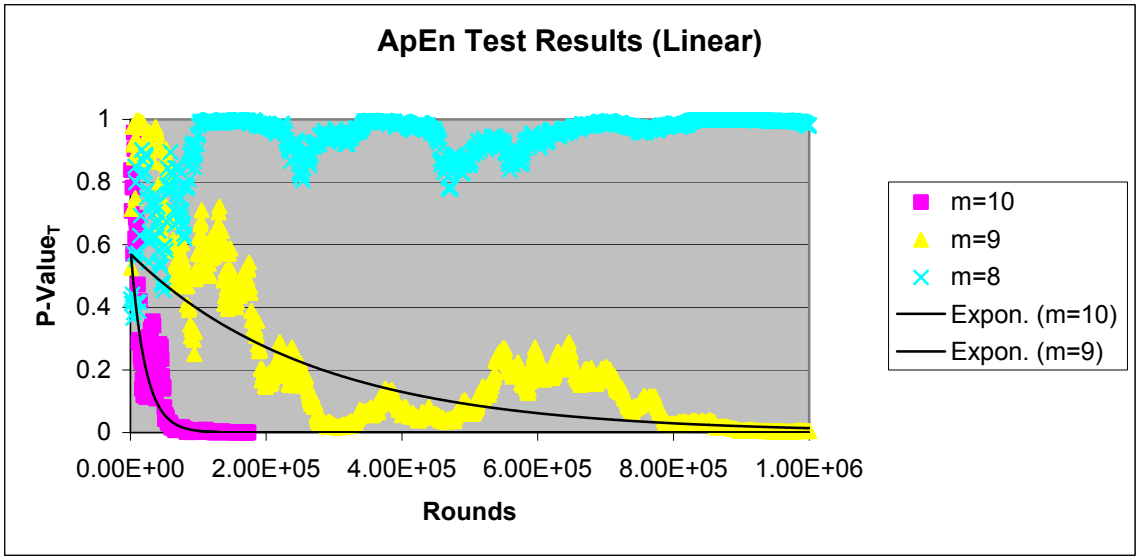

Figure 2: P-value$_T$ vs. number of rounds (linear)

This data suggests that the use of an ApEn block size of 10 for bit sequences of 1,000,000 bits is acceptable for up to 10,000 rounds of testing. This parameter selection becomes marginal if testing has to continue in order to resolve early failures; one could expect parameter-induced failures with this selection after approximately 100,000 rounds of testing.

Similarly, an ApEn block size selection of 9 (once again for n=1,000,000) would be acceptable for up to 300,000 rounds of testing (and thus an ApEn block size selection of 9 would be acceptable for most styles of testing). This parameter selection is also marginal for rigorous testing because of the general downward trend in P-value$_T$.

The ApEn block size selection of 8 (for 1,000,000 bit long bit sequences) appears to be acceptable for well over 1,000,000 rounds of testing, which makes this parameter selection reasonable for all practical testing strategies.

6 Conclusion

For sequences of 1,000,000 bits long, this testing reveals that the ApEn block sizes of 11-14 bits aren't useful for reasonable testing. The block sizes 13 and 14 consistently fail in less than 1,000 rounds of testing, and are thus not useful for any meaningful testing for this sequence length. The block sizes 11 and 12 are marginal even at 1,000 rounds of testing, and will not support even small amounts of expanded testing in order to resolve any testing issues; this restriction is unacceptable for serious testing.

For this sequence size, the ApEn block sizes 9 and 10 are acceptable for limited testing, as long as it is constrained (less than 10,000 rounds for the initial test set), and large-scale expansion of testing is not anticipated. In particular, testing with an ApEn block size of 10 becomes marginal after 100,000 rounds of testing, so any testing to and beyond this point is problematic.

Finally, for this sequence size, an ApEn block size of 8 is acceptable for all anticipated styles of testing.

7 Further work

Several tests would benefit from greater analysis of the acceptable ranges of parameter selection. Most of the parameter ranges have been established through asymptotic analysis, which doesn't necessarily apply to more constrained (i.e., non-infinite) testing. Empirical testing of each test for a particular bit sequence size (e.g. n=1,000,000) would be useful.

Additional testing regarding the selection of the ApEn parameter would also be useful. This analysis was made under the assumption that the FIPS 186-2 PRNG produced ideal output for the seed that was used. Additional testing could be accomplished by fixing a block size and number of rounds (e.g. ApEn block size of 10, with 100,000 rounds), and then performing a statistically significant number of these test sets (>> 10,000) with different initial seeds. If the proportion of these tests that passed was within the expected range, this would further support that this combination of parameters was valid. Unfortunately, this style of testing is not feasible to conduct with the available resources.

References

[0] Kim, Song-Ju, Umeno, Ken and Hasegawa, Akio, "Corrections of the NIST Statistical Test Suite for Randomness".

[1] Pincus, Steve and Singer, Burton, "Randomness and Degrees of Irregularity", Proceedings of the National Academy of Sciences, Vol 93, pp 2083-2088, March 1996.

[2] Pincus, Steve and Kalman, Rudolf, "Not All (Possibly) "Random" Sequences are Created Equal", Proceedings of the National Academy of Sciences, Vol 94, pp 3513-3518, April 1997.

[3] Rukhin, Andrew L, "Approximate Entropy for Testing Randomness", Journal of Applied Probability, Vol 37, 2000.

[4] Rukhin, Andrew et al, NIST Special Publications 800-22, "A Statistical Test Suite for Random and Pseudorandom Number Generators for Cryptographic Applications".

[5] Soto, Juan and Bassham, Lawrence, "Randomness Testing of the Advanced Encryption Standard Finalist Candidates", March 28, 2000.

[6] FIPS PUB 186-2 (+ Change Notice 2001 October 5), "Digital Signature Standard (DSS)".