

# Re: Interpreting the test result on a RNG

---

*Source:* <http://www.derkeiler.com/Newsgroups/sci.crypt/2007-02/msg00374.html>

---

- *From:* "Cristiano" <cristiano.pi@xxxxxxxxxxx>
  - *Date:* Mon, 12 Feb 2007 20:50:43 +0100
- 

Peter Pearson wrote:

On Mon, 12 Feb 2007 13:21:25 +0100, Cristiano  
<cristiano.pi@xxxxxxxxxxx> wrote:

On page 100 (page 109 of the pdf file) of this NIST paper  
<http://csrc.nist.gov/publications/nistpubs/800-22/sp-800-22-051501.pdf>  
there is an interpretation method of the result obtained from a test  
of a random number generator.

The range of acceptable proportion of sequences that pass the test is  
determined using the confidence interval calculated using a normal  
distribution as an approximation to the binomial distribution.  
I'd like to implement that method using the binomial distribution  
because the number of the tested sequences ('m' in the paper) can be  
50 or 100.

I calculated the following table with m=50 and a=0.05:

F	Bin	CUM(Bin)
0	0.076945	0.923055
1	0.202487	0.720568
2	0.261101	0.459467
3	0.219875	0.239592
4	0.135975	0.103617
5	0.065841	0.037776
6	0.025990	0.011786
7	0.008598	0.003188

F is the number of the sequences which failed the test (p-value < a)  
Bin is  $P_p(n|N)$  in this link (the binomial distribution):  
<http://mathworld.wolfram.com/BinomialDistribution.html>  
CUM(bin) is the probability to see more than F failures.

My problem is: which column should I use? Or, in other words, the  
overall test is one-tailed (column 'CUM(Bin)') or two-tailed (column  
'Bin')? I think the latter.

## Re: Interpreting the test result on a RNG

You could defend either decision, but pay attention to what you mean by your choice:

If you choose a two-tailed test, you're announcing that you're willing to tolerate a certain probability of mistakenly rejecting a perfect random-number generator for having either too many failures or for having too few failures. This is, in general, a reasonable thing to say (very few failures is, in fact, suspicious);

I totally agree.

but note that in the table you give the smallest possible number of failures, zero, has a 7.7% chance of occurring with a perfect random source. So any two-tailed test you apply would have at least a 7.7% chance of false rejection, which is probably more than you want to tolerate.

You're right, I didn't think.

In contrast, if you choose a single-tailed test, you're saying that you'll only reject a candidate source for failing too many tests -- and that you'll accept a candidate source even if the number of failures is anomalously small. If you're working from the table you gave above, this would be the reasonable thing to do. However, even if you work from the bottom row of the table (i.e., saying you'll reject the candidate if there are more than 7 failures), you'll have a .3% chance of rejecting a perfect random source. That's far too high for many applications.

Note that you have chosen as your acceptance criterion the number of occurrences of  $p\text{-value} < \alpha$ , rather than some other test of the hypothesis that the  $p\text{-values}$  are uniformly distributed. This might be an impractical choice when the number of  $p\text{-values}$  is relatively small: it seems to me that SP-800-22 talks in terms of 1000 tests, rather than 50 or 100.

In the current implementation of my test, I already use the KS test (and its Anderson-Darling variant) to check whether the  $p\text{-values}$  are uniformly distributed, I also use the Pearson's chi-square test with  $\sqrt{n}$  bins. I usually test 50 sequences with 100 to 1000 different tests, like this:

```
Seq#1 Seq#2 Seq#3 ... Seq#50 KS  
FFT test .9281 .1982 .0291 .1922 .1234
```

Re: Interpreting the test result on a RNG

Re: Interpreting the test result on a RNG

Serial test .1235 .6453 .3677 .7773 .4567  
Permut. test ...  
... ..  
Autocorr. test .4654 .3737 .1927 ... .3937 .6789  
Overall KS .9876

For each row, I calculate the overall p-value with the KS test, so I have 1 p-value for the FFT, 1 p-value for the Serial test and so forth.

When there are many tests, I get for some of them a very small KS p-value and I don't know how many tests with 0.0001... or 0.00001... can be tolerated.

To check whether the KS p-values are good (uniformly distributed) I do an overall KS test over the KS p-values; if this overall p-values is good it means that I can tolerate the small p-values.

Usually it works, but it's not rare to see a big failure even with a very good generator. For that reason I tried the binomial distribution method.

Probably the most useful next step is to decide what false-rejection rate you can tolerate. That would help to focus further discussion.

Not easy to say. I just need an objective method to say that a generator is bad with x% probability to be wrong.

Thank you  
Cristiano

.