

## Re: issues with statistical test suite from <http://csrc.nist.gov/rng/>

**Source:** <http://www.derkeiler.com/Newsgroups/sci.crypt/2004-01/1715.html>

---

**From:** Mack ([macckone\\_at\\_a\\_nospamjunk123\\_ol.com](mailto:macckone_at_a_nospamjunk123_ol.com))

**Date:** 01/26/04

Date: Mon, 26 Jan 2004 07:38:29 GMT

On Sun, 25 Jan 2004 13:23:29 GMT, "Cristiano"  
<[cristiano.pi@NSquipo.it](mailto:cristiano.pi@NSquipo.it)> wrote:

>Mack wrote:

>> On Sat, 24 Jan 2004 16:28:27 GMT, "Cristiano"

>> <[cristiano.pi@NSquipo.it](mailto:cristiano.pi@NSquipo.it)> wrote:

>>

>>> Mack wrote:

>>>> On Thu, 22 Jan 2004 19:30:50 GMT, "Cristiano"

>>>> <[cristiano.pi@NSquipo.it](mailto:cristiano.pi@NSquipo.it)> wrote:

>>>>

>>>> skewness should be less than  $2 * ses$  which will vary by sample size.

>>>>  $ses = \sqrt{6/n}$ .

>>>>

>>> I'm not a mathematician, so I don't know if this rule really

>>> applies. Could you elaborate a bit, please?

>>>

>>

>>  $ses$  is the standard error of skewness. It is similar to standard

>> deviation. Although taking the values as having the same

>> meaning is probably a bad idea.

>

>I think so.

>

>

>>>>>> Also the FFT  $p$ -values are not skewed (usually I get skewness=0.1,

>>>>>> 0.2).

>>>>>>

>>>>>>

>>>>>> Are you using the sample mean or expected mean? For the  $1e5$  FFT I

>>>>>> never got a skewness below .4. For 1000 samples .2 would

>>>>>> definitely

>>>>>> be a significant skew ( $2 * ses = .15492$ ).

>>>>>>

>>>>>> Sure, you use that test in a bad way;  $n$  must be around  $1e6$  bit, do

>>>>>> you remember?

>>>> *Anyway, your question seems strange. You must use the sample mean,  
>>>> not the expected one.*

>>>>

>>>> *That is incorrect when you are examining a sample presumed to be  
>>>> from a specific distribution. That would measure skew with respect  
>>>> to the sample itself, not with respect to the expected distribution.*

>>>>

>>>> *1e4 x 1000*

>>>>

>--

>>> -----

>>>> *C1 C2 C3 C4 C5 C6 C7 C8 C9 C10 P-VALUE PROPORTION*

>>>> *STATISTICAL TEST*

>>>>

>--

>>> -----

>>>> *0 0 0 0 0 0 0 0 1000 0.000000 \* 1.0000*

>>>> *Lempel-Ziv*

>>>> *igamc underflow error occurs for Lempel-Ziv*

>>>>

>>>> *As an example use the LZ test with 1e4 where all samples went to*

>>>> *bucket ten since the case statement doesn't handle this case. If*

>>>> *you*

>>>> *use the sample mean then it has a skew of zero. ie. the mean is ten*

>>>> *and all samples go to bucket ten. This is obviously not what we*

>>>> *want the test to show. We are looking for a measure of how well*

>>>> *this conforms to the expected distribution, in this case the mean*

>>>> *should be 5.5 and the sample is badly skewed.*

>>>>

>>>> *This is the first time I hear about skewness used for the goodness*

>>>> *of fit!*

>>>> *To see "how well this conforms to the expected distribution" you*

>>>> *must not*

>>>> *use the skewness, you must use a proper test (KS test, chi-square or*

>>>> *my SL*

>>>> *test, if you like). Why don't use also the kurtosis, the median and*

>>>> *so*

>>>> *forth? This way you have everything but the stuff you need.*

>>>>

>>

>> *Skewness is a parameter of "goodness of fit" as is kurtosis and*

>> *median. Generally they are used for goodness of fit to a normal*

>> *curve but they can be used for other distributions as well.*

>

>*I disagree.*

>*You can use those parameters to see how much \*they\* are different from the*

>*expected ones, but you shouldn't use them to see how much a set of samples*

>*differs from the expected distribution.*

>*I mean that if you see skewness=.65 you are not able to say how much your*

>distribution differs from the expected one. Obviously you could say that  
>your distribution is not perfectly good, but how much?  
>On the contrary, with a proper test for the goodness of fit you are able to  
>calculate a p-value to say how much your distribution differs from the  
>expected one.

In this specific case the p-values from a KS test were showing a value less than .001, quite a bit different from uniform. The skewness is a measure of how it isn't uniform. In a normal distribution it would be a measure of how it differs from the normal.

>  
>For example, I use the FFT to transform 32 kbit and then I calculate the  
>mean and the skewness of the transformed values.  
>If I calculate those parameters for a good prng and for a bad one (like  
>lcg), I seen that they are very different.  
>Unfortunately I can't know if a generator under test is good or bad if it  
>gives values in the range skewness\_bad ... skewness\_good because I don't  
>have a significance level.  
>To calculate a significance level I could calculate KS test of the  
>transformed values, but I don't know how to do it.

The traditional skewness level is 2\*ses. Anything outside of this is definitely bad if it is recurring. Anything inside of 1/2\*ses is probably ok.

>  
>  
>> Skewness measures symmetry about a point. Kurtosis could be  
>> used but the expected value would not be zero as for the normal  
>> curve when applied to a uniform distribution, although this can be  
>> easily calculated.  
>  
>Sure, it is  $6/5 * (n^2+1) / (n^2-1)$  for a discrete uniform distribution.  
>And when you got, for example, 7/6 what do you say? It is good? It is bad?  
>And how much?  
>  
>Here I have two doubts:  
>1) Surely we get some information from those parameters, but can the  
>information gotten be used in testing a rng (in an efficient way)?  
>2) You say: "Skewness measures symmetry about a point". I don't know how you  
>calculate "your" skewness. Do you calculate it using the absolute moments or  
>the central moments in some "strange" way?

skewness= $1/n * (\text{sum from } 1 \text{ to } n((Y_i - \text{expected mean})^3)) / \text{expected deviation}^3$

specifically:

Skewness is a measure of symmetry, or more precisely, the lack of symmetry. A distribution, or data set, is symmetric if it looks the

same to the left and right of the center point

from: <http://www.itl.nist.gov/div898/handbook/eda/section3/eda35b.htm>

and for Kurtosis:

Kurtosis is a measure of whether the data are peaked or flat relative to a normal distribution. That is, data sets with high kurtosis tend to have a distinct peak near the mean, decline rather rapidly, and have heavy tails. Data sets with low kurtosis tend to have a flat top near the mean rather than a sharp peak. A uniform distribution would be the extreme case.

>  
>  
>> *We have already agreed that KS is not the right test here. Chi-square or SL are more appropriate.*  
>  
>Yes, or perhaps no (see next paragraph).  
>  
>  
>>>> *Diehard doesn't give KS results except where it is appropriate.*  
>>>  
>>> *So does NIST test. But exactly, what do you mean?*  
>>  
>> *The finalAnalysisReport returns KS test values where these values are not appropriate.*  
>  
>Who say that? Have you done a new discovery?  
>If you calculate the KS test for \*only\* one sequence, then the KS is good enough.

Why would you ever use a KS test on one p-value?

>But if you calculate the KS of the KS's gotten from 100 or 1000 sequences, then the overall p-value is useless because the 100 or 1000 p-values are too binned.  
>  
>  
>>>> *Unfortunately the output is pretty hard to read.*  
>>>> *I usually open it with a text editor and search for results of .000, .00, and .0.*  
>>>  
>>> *And when you find them what do you do?*  
>>  
>> *Repeat that specific test with more data to determine if it is isolated or consistent.*  
>  
>With more data? Each test needs a fixed number of 32-bit numbers (some test requires slight variations on the number of input numbers).

>  
>Anyway, when a generator is definitely good or bad?  
>

A generator is bad when it fails a test that has a good mathematical foundation in a spectacular manner Or alternatively you can say that a generator is bad when fails some test (mathematical or empirical) that other "good" generators pass. Of course a single number that is .0000 or .9999 is not an indication of failure it must do so consistently, since a value like that happens randomly 1 time in 10000.

A generator is good when it meets the criteria for which it will be used. A bad generator such as a simple congruential generator may be a "good" generator if we are using it in a non-demanding application.

>  
>> *The newer version of diehard returns*  
>> *a final KS value but also states that it is more of a general*  
>> *guide than absolute result.*  
>  
> *That final KS p-values seems really useless calculated that way, because it*  
> *is the p-values of heterogeneous p-values.*  
> *I found much more useful to calculate the overall p-value for each test done*  
> *100 times; if I have 16 tests, I get 16 overall p-values and then you can*  
> *see where's the problem.*  
>  
>  
>>>> *I am also having to create my own test suite because nothing*  
>>>> *else meets my current needs. sts seems like a good package but*  
>>>> *it has its limitations.*  
>>>>  
>>> *Yes, all the tests have limitations. I think if one uses a test in a*  
>>> *proper*  
>>> *way the test can be useful anyway. The "proper way" could be also to*  
>>> *discard*  
>>> *a test! I done that with some test in dh.*  
>>  
>> *I have never found it necessary to discard a DH test. They may not*  
>> *detect a problem where there isn't one but they have never given*  
>> *a strong result of a problem where one didn't exist.*  
>  
> *I don't know the status of the newer version of dh, but that test has had*  
> *many problems (for example you could see my post on september 2003 about the*  
> *bad distribution of the overlap sum test).*  
>

The newer version is still being listed as 0.2 beta. However that error has been corrected. The big three tests Gorilla, GCD, and Birthday Spacings test all seem to be functioning adequately.

sci.crypt: Re: issues with statistical test suite from <http://csrc.nist.gov/rng/>

>  
>> *I am still a bit suspicious of the FFT and LZ tests since they do not*  
>> *yet have a firm mathematical foundation. They seem like good tests*  
>> *but they are still empirical. Of course we should be suspicious of*  
>> *any single test only by using a number of tests can we be certain that*  
>> *we aren't getting false positives or negatives.*  
>  
>*I'm used to testing the test to avoid some surprise.*  
>  
>*Cristiano*  
>

Leslie 'Mack' McBride  
remove text between \_ marks to respond via e-mail