IV THE TESTING PROGRAM FOR ACCURACY

WHAT ARE THE ISSUES

This is a broad area and it involves several questions.

1. How can we determine that a software program works, and gives correct values?
2. What is accuracy and precision of a statistical software program?
3. Can we depend on the program to always give accurate results?
4. What are the standards, what are the tests, what are the measures?
5. Can we predict the accuracy of functions and routines?

SOFTWARE PROGRAM TESTING

This is essentially to answer issue 1.

"Any testing of statistical software programs involves the exercise of selection to get down to the area or routines to be tested. With respect to Excel these are functions and data analysis routines".

Since Excel is a commercial fully developed software program, the only user testing that can be done is to determine performance, accuracy and allowable input variable ranges. The only observable, measurable objects that can be evaluated are the numbers (or error flags) that come from the Excel function or Data Analysis routine. There are no internal values that are accessible.

The issue of hardware and compiler differences (Colonna 1993) as a cause of numerical differences, is not important here. The Excel package is a compiled version, and is primarily used in computers that recognize the Intel instruction set.

"For accuracy testing the software will require a test database and a parameter and selection vector. In some cases only a test database is needed and in some others such as the distribution functions, only a parameter vector is needed. In all cases there has to be an output vector that can be compared to a reference standard vector, such that a difference can be obtained as a measure of the accuracy of the method. In the case of Excel functions, this output vector has only one value (the exception is the array functions that output a range, matrix or a table of values). The Excel Data Analysis routines also may output a table, which is the output vector formatted to be readable.

In general, there are four methods to assess the accuracy of the output numbers:

> Method A: “Theoretical values manually calculated or selected (by theory) of the test data base that are valid accurate reference values. For example one can construct a list of data values that has a theoretical precise mean and a precise standard deviation.
>
> Method B: Values calculated by an external software program, chosen to be the reference or standard.
>
> Method C: Values published as part of a standard.
>
> Method D: Comparisons of outputs among two or more different software programs that have a common input data set.

For the larger, complex software programs, testing for validity and correctness has been a reported problem since The American Statistician (TAS) started “Statistical Computing Software Reviews and Editor’s Notes”. Author’s writing reviews have repeatedly commented about the lack of standards with which to evaluate software programs.

Method D has a different set of problems, 1) there are different assumptions inherent in each software package that are usually not “published” 2) there are different algorithms and different computing sequences which do affect output values and 3) the input commands and controls may not be completely described by the vendor, or may not be adequately understood by the tester. They also may not affect the same parts of a complex algorithm. However method D is in actuality, the only tool available today to access and test statistical software.

The Editor’s Notes (Goldstein 1994) illustrates the problems in trying to understand the differences between SAS and SPSS outputs for Dallal’s data from a 2 X 3 two-factor experiment. Delong’s comments (DeLong 1994) on the basic problem of determining the appropriate statistical tests on data from an unbalanced design. (i.e. the different software programs may be doing different tests in different ways.)

Methods A and C have been the historical ways to evaluate Excel. The NIST data sets and precise statistical values, which NIST reports, represents method C, and it has been the primary basis for determining the accuracy of Excel.

ACCURACY AND PRECISION

This is to try and resolve issue 2.

WHAT DO THEY MEAN?

“Precision concerns the tightness of a specification. Accuracy concerns the correctness of the specification An utterly-inaccurate-statement can be uttered quite precisely. For example, 3.17777777777777 is a precise number, but is an inaccurate approximation to Pi. Precision is to accuracy as intent is to accomplishment.

“A number possesses only its value. A number can represent only itself, and does that perfectly. By themselves, numbers possess neither precision nor accuracy.

"Precision attaches to the format into which the number is written or stored or rounded. Precision tends to be identified with the number of digits displayed. However a one digit integer is considered more precise than a floating point number. The term 'Relative Precision' refers to the number of significant digits, regardless of the significance of the number." (Kahan 2004)

Kahan here does not address the primary problem of language and units of measure. If for example the measure is probability, then a p value of 0.00003198 has a precision of 8. If however the measure of length is meters, then a measurement on the size of a biological cell would be a very small number, a number preceded by 8 zeros, none of which bear any relationship to the precision of the optical measurement of biological cell. In many cases we use particular units of measure such that measurements can be clustered about a decimal point. Representing the measurement in floating point format however leads to a more convenient way of identifying precision. Precision can then be ambiguous unless it is connected to a defined situation of measurement.

"Accuracy connects a number to the context in which it is used. Without its context, accuracy makes no more sense then the sentence 'Rosco is very tall' does before we know whether Rosco is a building, an elephant, a sailboat, a pygmy, or something else. In context, higher accuracy implies smaller error. Absolute error is the difference between the number you got, and the number you desired.

"Today's common programming languages pose no insurmountable obstacle to satisfactory floating point accuracy. It is limited by a programmer's cleverness and time.

"The precision declared for storing a floating point variable, the accuracy with its value approximates some ideal, the precision of arithmetic performed subsequently upon it, and the accuracy of the final result computed from that value cannot be correlated reliably using only the rules of a programming language without error-analysis." (Kahan 2004)

WHAT IS THE SPECIFICATION HERE?

The fundamental issue becomes one on how should we evaluate a string of decimal digits representing a number as a computer output?

There are essentially two bases for identifying a specification (in Kahan's view).

BASIS ONE

The first basis is the representation of a number, based on the number being a "textural string of digits." This is the situation in which the number of digits describing the number is unlimited. The true value of pi is in this category, where the string can contain an infinite number of symbolic digits, representing a number system to the base 10. Other bases such as octal hexadecimal or binary can also be considered. Any finite string then is always an inaccurate representation of pi. If we take pi as our specification, then we can state a precision of its representation as being a string say of 24 decimal digits. Note, that 24 digits is an arbitrary number of digits. We have no apriori rational for selecting 10, 15, 24, 30 or any number of digits. With the precision then set to 24 digits, then we have a basis for evaluating the accuracy of any algorithm that calculates pi.

BASIS TWO

The second basis is the concept of numbers that represent a form of “measurement”. Measurement is always finite and carries other implications that Kahan talks about, such as integer measurements “counts of”. Kahan’s arguments are based primarily on the idea that the numbers represent “measurements”. Computations then on these measurements need only the precision of the basic measurements.

THE COMBINATION OF BASES, A PROBLEM

From a statistical standpoint, the first basis is the realm of statistical distributions and theoretical mathematical constructs and the second is data. When these are combined, we now have problems in deciding on what the precision is of the result should be. A precise and accurate combined computation to 100 digits involving data accurate to only 3 significant digits, is “an overkill”. However we can certainly state that the statistical computations must not degrade the accuracy of the input data. Therefore we don’t require 100 digit accuracies when we only have data good to 3 decimal places, but we do need something better than 3 decimal places.

THE FIXED DISPLAY, ANOTHER PROBLEM

This occurs when the output is “locked” into a fixed number of digits. This is not a problem with Excel, since the display is always changeable by means of Cell Format settings. It is however a major problem with free statistical software and with some commercial software such as Minitab. There are several internet sites that list free statistical software. These all generally have fixed display routines with 3 or 5 decimal digits displayed. The basis then for testing is to assign all of the trailing digits to zero. Consequently 3 digit displays will always have poor (or failing) test results.

MEASURES

ACCURACY OF A COMPUTED NUMBER

Accuracy is basically comparing a function/routine output (CV) on a given data set to a known measure of the data set (RV). RV values could come from precise computational values or from artificial data sets that have theoretical, precise or preset measures (such as standard deviation), The NIST testing approach described below uses both methods

LRE MEASURES

The measure of the accuracy of the information from a computed value is by a calculation called Log Relative Error or LRE. This was introduced by McCullough (1998) in his article on testing statistical software. The LRE value represents a measure of how many significant (accurate) digits (decimal) there are in the output parameter values.

1) **LRE = $-\mathrm{LOG}_{10}$ (abs (CV-RV) / RV))**

Where: CV is the computed value

RV is the reference or true value

Under the general accuracy limitations, the maximum LRE value for double precision values is 15.7. This is where the calculated value and the reference value are identical (bit for bit for the entire 64 bit floating point number). The agreed on accuracy in conversion from binary to decimal is an accuracy of 2E-16, which is equivalent to an LRE of 15.7. In Excel, the maximum number of floating point digits that can be displayed is 15. A one unit difference between numbers in the 15th digit, involves a difference in the least 4 bits of the 52 bit mantissa, and can result in LRE values as low as 13.5, because the automatic rounding of decimal digits also has an effect. McCullough (1998, 1999a) and others consider 15 as the maximum LRE value. As McCullough (2003) has stated, an LRE value of 15 refers to the fact that the calculated value agrees with the 15-digit (decimal) reference value given with the data set, digit for digit. For that reason, we can consider LRE values of 15 or higher as indicating an exact match to the reference value.

When the reference or true value (RV) is zero the above equation will not apply. In this case,

2) LRE = -LOG_{10} (abs (CV))

Where: CV is the computed value is used.

LAE MEASURE

For probability values, a relative reference may not be a good baseline, and an absolute measure called the Log Absolute Error or LAE (McCullough's LAR, 1998) is often used.

3) LAE = -LOG_{10} (abs (CV-RV))

Where: CV is the computed value

RV is the reference or true value

It is a measure of how many significant (accurate) digits there are in the p value to the right of the decimal point, including zeros. In most cases, an accurate p value within .000001 to .999999 is adequate, which correspond to LRE values of 6 or more. Again the scale is from 0 to 16.

DATA SET MEASURES

It is convenient to show graphs of changes in computation accuracies on a given data set due to a change in data set characteristics. Therefore the following measures represent many possible data set characteristics that could be used to show how accuracy changes as these characteristics change.

SIZE

This is the number of cells in the range of the data.

MAGNITUDE

There are different interpretations of "magnitude". In the floating point representation of numbers, magnitude is usually taken as the numerical value of the exponent (the nn number), which can vary from positive (large magnitude) to negative (small magnitude).

NUMBER OF SIGNIFICANT DIGITS (NSD)

THE CORE NUMBER

If we take a floating point number in a cell, and represent it as a string of decimal digits without an exponent, the string would have a decimal point somewhere in the string, and would have a contiguous group of digits that are (conceptually) preceded by zeros and trailed by zeros. This group (the core number) represents the measurement without any magnitude representations. Accuracy is focused on the core number, not on magnitude because of the floating point number base.

THE FLOATING POINT NUMBER

When magnitude representations are added in, the number in a cell is displayed in a "scientific" format. The group may include internal zeros as part of the measurement, and may also include or exclude the decimal point.

When the number is represented by a floating-point scheme, the core number could be completely represented by the numbers left of the "E" symbol and the magnitude by the number right of the "E" symbol.

NUMBER OF SIGNIFICANT DIGITS (NSD)

The number of significant digits is the number of decimal digits in the core number. It represents the measurement without any magnitude or scaling information. We often add zeros to the right of a core number for convenience of display, but these added zeros do not increase the number of significant digits.

4) NSD = Number of significant digits in core number

The core number can however include zeros on the right as significant digits if they are part of the measurement.

A typical set of actual measurements or data would have core numbers having different numbers of significant digits.

CELL-TO-CELL VARIATION

These are the cell-to-cell number changes that are the main focus of statistical analysis. They represent the changes in measurements due to variable manipulations and random effects. For the test data sets used to test Excel, the cell-to-cell variation is primarily the number changes in the digits to the right of the decimal point.

The standard deviation is a good measure of this el-to-cell variation

COEFFICIENT OF VARIATION (COV)

This is a general measure of the test data set.

5) COV = (Standard deviation, cell to cell) / (Average of all data set cells)

OTHER MEASURES

RE-SCALING

A given data set represented as a set of floating point numbers, can be re-scaled without changing the internal, patterns within the data set. The re-scaling changes the virtual location of the decimal points. By appropriate choice of the multiplier/divisor as a re-scaling factor in the form of 10^m, where m can be positive or negative, the decimal points can be moved right or left. Given the floating point form of the number (in decimals), the rescaling does not change the characteristics of the number set.

A re-scaled data set is one where we now have digits to the left of the virtual decimal point as being unchanged from cell-to-cell, and the digits to the right of the virtual decimal point as being all of the values that change.

This particular form works only when there are only a few digits that change on the right of the decimal point.

NCD MEASURES

This is the number of significant digits on the left of the decimal point when the cell-to-cell variations are all on the right side of the decimal point. These are the number of digits that do not change from cell-to-cell within a designated range.

6) NCD = Number of digits left of decimal point after re-scaling

It is a representation of the magnitude of a rescaled number. It is restricted to positive values. It also applies to situations where the changes on the right of the decimal point do not involve more than 2-3 digits.

NCD as a measure of a data set is the same, whether the data set is in the original version or in the rescaled version. The rescaling allows us to correctly define the NCD value. It also allows us to manipulate the rescaled data set to test for computational inaccuracies.

LIC MEASURES

Some test data sets can’t be rescaled to get LCD values, because the variations involve more then 2 decimal digits. The effect measure then comes from the COV value.

7) LIC = ABS(-LOG10(COV))

Where COV is as defined above as 5)

LTP MEASURES

This is a logarithmic measure of a p or q value either output from or input to a probability function. It is a convenient plotting measure, defined as follows.

8) For p values less than or equal to 0.5, LTP = $\log_{10}$ (p)

8) For p values greater than or equal to 0.5, LTP = - $\log_{10}$ (1 - p) or

8) LTP = - $\log_{10}$ (q), if the q comes from a different function.

There is a gap between –0.3 and +0.3 where there is no LTP value. The LTP = 0 point represents a p value of 0.5. To close the gap would require unequal x axis intervals, and Excel charts can’t express this. In some charts, a simple horizontal line from –0.3 to +0.3 is used to show and fill this gap.

The log(p) value then retains the information content, and allows the use as a generalization. Graphics are considerably improved when LTP values are used for p or q values instead of scientific notation.

Since the smallest p value that can be obtained in Excel is rmin, expressing the minimum floating point number, as a LTP value becomes –308.34734. Consequently a chart with a horizontal axis scale of –310 to 0 covers all left tail p values, and right tail values can be covered with a horizontal scale of 0 to 15.

DIFFERENCES BETWEEN NCD AND LIC MEASURES

NCD values take on integer values, while LIC values are continuous. Therefore they are not exactly equal. LIC values are easier to compute, since there is no rescaling needed.

When the number of constant digits gets over 10, LIC values may become irratic due to the inherent problems with the Excel STDEV function. At this point one needs to shift to NCD values, since they are not dependent on any Excel function or routine.

Both NCD and LIC are measures of the number of significant digits.

For the original Simonoff data (Table 4-1 for Y), rescaling gives NCD = 9 and the variations of table 4-2. The standard deviation in table 4-2 is 0.05278 and the average is 0.061. Combining for table 4-1 (Y) we get a COV value of (0.5278/(1E+10 + 0.61)) = 5.278E-10, and a LIC value of 9.28.

THE DATA SETS FOR TESTING

THE NIST DATA SETS

One of the common sources for testing the performance of statistical software is the StRD data sets found on the Internet at the National Institute of Science and Technology (NIST) web site. Publication #1396 from NIST gives further information on the StRD set. There are four categories of data sets, with separate data sets within each category for a total of 31 data sets. They are Univariate, One-way ANOVA, Linear Regression and Non-Linear Regression data sets. These sets are used to determine (a), the ability of a given program to handle the data set, and (b), to determine the accuracy of output parameter values and (c) to reveal something about the underlying algorithm (McCullough 1999). The NIST reference fully describes each data set, and how precise parameter values were obtained. However the NIST suite is very limited and scientifically oriented, and does not fully test a software package for accuracy.

The NIST suite tests a minimum core capability of statistical software. It does not have any extended data sets to test the much larger capabilities of current statistical software packages. Reviewers of commercial software have complained about this limited capability. New software coming out, and the expansions in older packages address newer statistical methods and newer statistical tests that are beyond the scope of the NIST suite.

The NIST data sets are in some cases, real data (measurements) and in other cases, contrived data having a special characteristic. This special characteristic consists of a set of numbers having variable digits to the right of a decimal point and constant digits to the left. The constant digits are the same within a given data set.

For example a large data set would be alternating decimal values of n.3 and n.1, where n would be 1E+03 for one set, 1E+06 for another set, 1E+09 for another set and 1E+12 for still another set. The characteristics of these contrived sets are that they have a precise, low known variance and a precise mean. This approach to test software programs was established many years ago, and is biased to show performance different from what would occur with more realistic data sets.

THE ISSUE DATA SETS

These are a number of data sets in e-mails, websites, etc that were specific to faults and errors in Excel. They were presented to point out a particular problem with Excel/ Most of them came out after Excel 2000 was released. Table 4-1 is one of these.

Table 4-1: The Simonoff Set On LINEST

Row	X	Y
1	10000000001	1000000000.000
2	10000000002	1000000000.000
3	10000000003	1000000000.900
4	10000000004	1000000001.100
5	10000000005	1000000001.010
6	10000000006	1000000000.990
7	10000000007	1000000001.100
8	10000000008	1000000000.999
9	10000000009	1000000000.000
10	10000000010	1000000000.001

Other sets are described within the section where they are tested or referenced

THE OFFSET DATA SERIES

These are sets of special data sets intended to test the function/routine for accuracy when the means change and the size and cell-to-cell variation held constant. The data set is in a re-scaled form, allowing constants to be added to the numbers without changing the basic variations within the data. For example, the NIST StRD NumAcc series include 2 individual offset data sets. These essentially test the effect of a change in the LTP values with the variations remaining the same from set to set. For this illustration, the NumAcc sets were expanded from 2 to 16 sets, covering magnitudes from 0 to 10^16 (LTP values from 0 to 16).

CENTERED DATA SETS

One method used was to take a known data set and center all the data about the respective mean, in double precision. Then the routine/function being tested was run on the centered data to create a set of reference values. The routine/function was then run on the original data, and LRE values calculated on the difference between the two results. If the LRE values of the difference between the two were in the range of 16 to 12, the centered data generated reference values were then taken as a “correct” reference set.

THE INADEQUACY OF TEST DATA AND ACCEPTED STANDARDS

This is the lack of adequate test data sets in the literature.

Currently, the NIST STRD suite is considered inadequate to test current new statistical software programs and expansions of older packages such as SAS, SPSS, etc. There are a very large number of statistical applications/tests in commercial software these days for which there are no accepted test sets to evaluate performance. As a result, current software reviews have degenerated to just descriptions on what they do. (See the issues of The American Statistician from Feb 2000 to Aug 2005. 15 software reviews/tests were conducted, 1 using the NIST suite and 5 using some tests developed by the author of the review. Out of the 15, 14 were beyond the scope of the NIST suite.)

GENERAL RELATIONSHIPS BETWEEN LTP AND LRE VALUES.

THEORETICAL BASIS

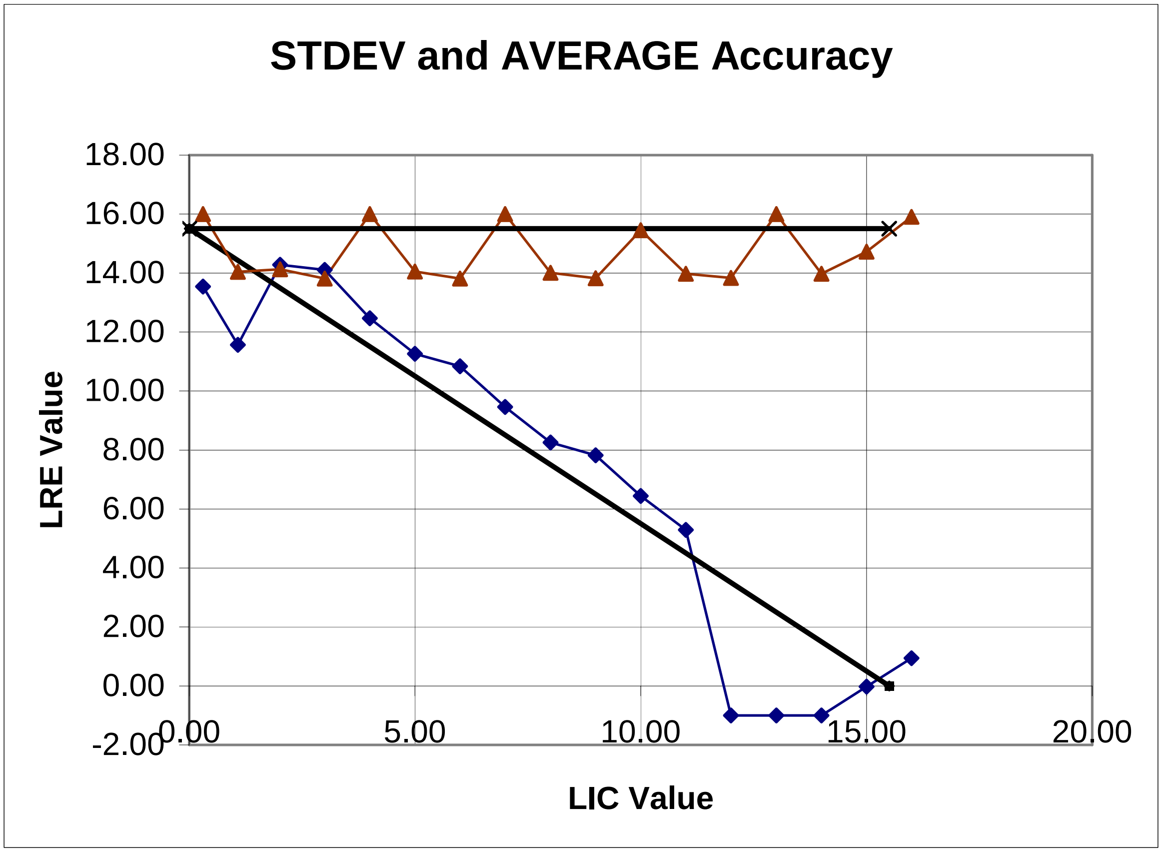
Given the fact that Excel computations conform to the IEEE-754 standard (see Section 3), there is a general consistent relationship between data set LIC values and function LRE values.

- When the function/equation/routine/macro **only** does multiplications and/or divisions on numbers, there is no change in LRE values when LIC values change.
- When the function/equation/routine/macro includes addition on numbers, then there is a slight decrease in LRE values as LIC values increase. This is due to difficulty of doing accurate additions, as discussed in section 2. The LRE value is sequence and magnitude dependent. The effect cannot be predicted.
- When the function/equation/routine/macro includes subtraction, then there is a big change in LRE values. This effect can be predicted. The relationship is:

  LRE = -LTP (or LIC) + 15.75 Eq. 1

These effects are shown in Figure 4-1 in relationship to the NIST extended NumAcc data set of floating point numbers.

Figure 4-1: Theoretical IEEE-754 (Double) LRE Characteristic With Data



The AVERAGE function only does summations and division. It is quite uniform because the data is uniform in magnitude within each set. The STDEV function does addition, subtraction, multiplication and division.

LINEST also shows these relationships.

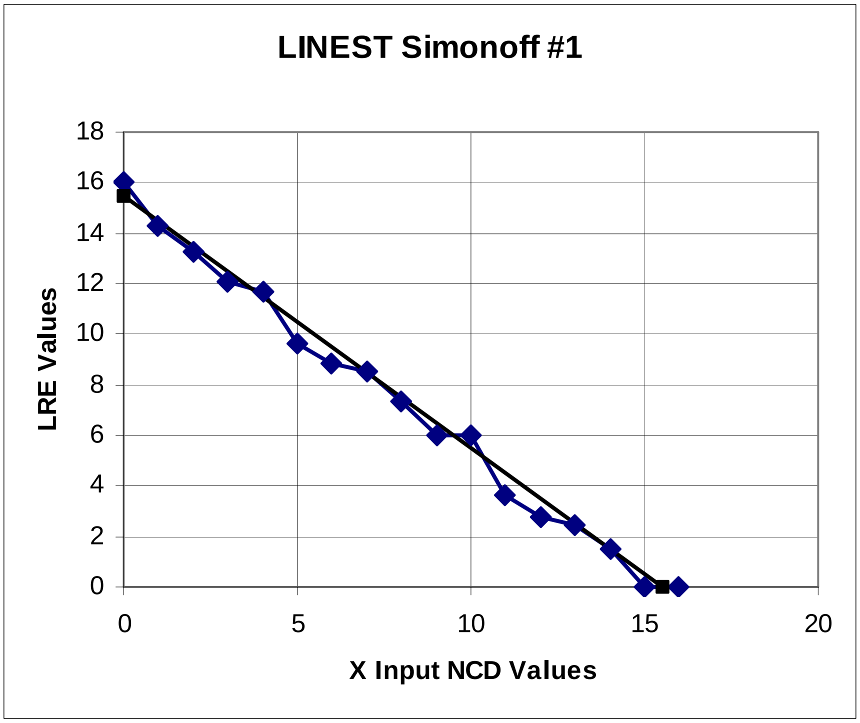
Take the Simonoff, table 1 on LINEST from above, remove the constant digits to give the basic variations.

Table 4-2: Data from Table 4-1 Reduced to Variations Only.

X	Y
1	0.0000
2	0.0000
3	0.0900
4	0.1100
5	0.1010
6	0.0990
7	0.1100
8	0.0999
9	0.0000
10	0.0001

Now by keep adding a constant to the Y values and doing a LINEST of the new X and Y pairs, we get the following chart of LRE errprs

Figure 4-2: LINEST Accuracy, Simonoff #1 Data



The straight black line represents equation 1. We can conclude then that the theoretical basis and the accuracy measure discussed above seems to apply broadly. LTP/LIC is a valid measure of magnitude.

From this we can conclude:

1. There is a limit on the accuracies of computation, and that limit is inherent in the IEEE-754 floating point number standard.
2. Accuracy of a computational sequence can be improved in some cases by subtracting the common constant from the original data set. The error in subtracting floating point numbers from a constant is still there. By using the mean, the subtraction error magnitudes are smaller.
3. The accuracy of a computation by one of Excel's statistical functions/routines with this data set can be approximately predicted from figure 4-1.

ACTUAL ACCURACIES, SOME GENERALIZATIONS

Some generalizations can be made about the actual accuracies of computations, based on observations. Here the observations are from Excel 2003 computations using the Intel Pentium 4 CPU at 2.80 Ghz, operating under the Windows XP environment. Users of Excel 2003 on HP or RISC type computers that do not support the IEEE-754 Long Double, can expect lower LRE values. A RISC type computer was not available, so no quantitative assessment of the loss in accuracy can be made.

When there are involved computations such as fitting equations to data, regressions, large matrix manipulations, etc., the accuracy of results as LRE values generally runs from 9 to 12. During general computations, LRE values can run above 12, into the 13-14 range. In a general sense then we can say that actually the displayed 15 digits are accurate to about 12 digits. This is due to the basic limitations of a computer doing many sequentially related computations using IEEE-754 double precision numbers. The problems here are discussed in section 3. Summations of series (such as used for statistical distributions) generally have LRE values that fall in the 9-12 range, but there are times when the values have lower accuracies, and get into the 3-6 ranges. The different sections in this report describe situations where low accuracies (0 to 3) actually occur.

Excel accuracy could be increased by full use of the long double floating point (use of 80 bit registers and memories to hold intermediate computations). Kahan (2004) argues for this design If all floating point values were held as long-doubles, this would put the maximum LRE value to 19.5, which is about a 4 digit gain in accuracy over the standard Excel 15 digits

WHAT IS THE ACCEPTANCE CRITERIA

THE VIEWS OF ACCEPTANCE / REJECTION

1. Every computation in Excel should be accurate to 100 decimal digits (i.e. a non-standard floating point form with a mantissa of 100 digits), There are critics that take this view, and totally dismiss Excel. However most commercial software does not provide 100 decimal digits, but it however is acceptable to these critics.

2. Every computation in Excel should be accurate to 14 floating point decimal digits. This is within the IEEE-754 floating point standard. This would be a possible goal, but would be very difficult to fully meet for ALL POSSIBLE DATA INPUTS.

3. Accuracy as a IEEE-754 limit is acceptable. The facts are that the theoretical line of figure 4-1 is not realized by some Excel functions over the entire LTP range of 0 to 15. An example is the Excel STDEV and VAR functions. The actual behavior on test data sets is not even. For LTP values of about 12 or more, the functions give totally wrong values (LRE’s of zero). The issue then is, do we accept this, because this type of data would be very rarely encountered, or do we reject Excel because of the observed failure to follow the IEEE-754 characteristic.

For all the critics taking positions 1 and 2 should obviously use some other software program, and dismiss Excel.

The ‘ views are that 3 is acceptable, but we should push Microsoft to improve Excel software algorithms to where we can say then that Excel accuracy is only limited by the IEEE-754 double standard.