

V. UNIVARIATE ANALYSIS .....	2
UNIVARIATE FUNCTIONS: .....	2
UNIVARIATE DATA ANALYSIS – DATA ANALYSIS ROUTINES: .....	4
FUNCTIONS/OUTPUTS PROVIDED.....	4
NOTES AND COMMENTS ON THE PROVIDED FUNCTIONS / OUTPUTS .....	5
VARIANCE.....	5
STANDARD DEVIATION.....	5
SKEWNESS AND KURTOSIS .....	6
CONFIDENCE LEVEL.....	8
REPORTED PROBLEMS:.....	9
RESULTS OF ACCURACY TESTS:.....	9
DIFFERENCES BETWEEN EXCEL VERSIONS.....	9
TEST GROUP 1: .....	10
TESTS CONDUCTED.....	10
TEST GROUP 2: .....	12
TESTS CONDUCTED .....	12
TEST RESULTS.....	12
TEST GROUP 3: .....	13
TESTS CONDUCTED .....	13
TEST RESULTS.....	14
TEST GROUP 4: .....	18
COMMENTS, FIXES, WORKAROUNDS AND RECOMMENDATIONS .....	20
PROBLEMS WITH VARIANCE CLASS FUNCTIONS, EXCEL 97 and 2000 .....	20
OVERVIEW AND COMMENTS.....	20
FIXES, WORKAROUNDS AND RECOMMENDATIONS: .....	20
PROBLEMS WITH VARIANCE CLASS FUNCTIONS, EXCEL 2003 AND 2007.....	20
OVERVIEW AND COMMENTS.....	20
FIXES, WORKAROUNDS AND RECOMMENDATIONS: .....	20
PROBLEMS WITH THE GEOMETRIC MEAN EXCEL 2000 AND 2003.....	21
OVERVIEW AND COMMENTS.....	21
FIXES, WORKAROUNDS AND RECOMMENDATIONS: .....	21
PROBLEMS WITH RANK AND PERCENTILE .....	21

OVERVIEW AND COMMENTS .....	21
FIXES, WORKAROUNDS AND RECOMMENDATIONS: .....	21
PROBLEMS WITH AUTOCORRELATION.....	21
OVERVIEW AND COMMENTS .....	21
FIXES, WORKAROUNDS AND RECOMMENDATIONS: .....	22
CONCLUSIONS.....	22

## **V. UNIVARIATE ANALYSIS**

### **UNIVARIATE FUNCTIONS:**

These are operations on one-dimensional lists. They provide a statistic on data in a list. The statistic characterizes some property of interest that the data list has. In Excel they are calculations from functions or the results of a routine from the Data Analysis Tool-pack. The list being analyzed can be a column or row of data, or an entire block (rows and columns) of data.

There is no uniform accepted standard list of names of terms or equations that represent a standard to which a software program is required to provide values for.

In Excel, The following functions are considered univariate functions. They are entered into a cell as a formula with the input as a range of cells. They only operate on cells with numbers in the input range.

**AVEDEV**  
**AVERAGE....(see table 5-1 variations)**  
**COUNT (1, 2)**  
**DEVSQ**  
**GEOMEAN**  
**HARMEAN**  
**KURT**  
**LARGE**  
**MAX.....(see table 5-1 variations)**  
**MEDIAN**  
**MIN..... (see table 5-1 variations)**  
**MODE**  
**PERCENTILE**  
**PERCENTRANK**  
**QUARTILE**  
**RANK**  
**SKEW**  
**SMALL**  
**STDEV.....(see table 5-1 variations)**  
**TRIMEAN**  
**VAR.....(see table 5-1 variations)**  
**CONFIDENCE**

Some of these have variations. By use of preceding and trailing letters, the way the function interprets inputs and acts on the data, changes. For example, DVAR, VARP, VARA and VARPA, all refer to the basic variance computation, but have different relationships to the range designation, to the way non-numbers are treated, and to the sample/population calculation. For example VARP means that only numbers are used, and the range of the data is treated as a population.

**Table 5-1: Function Naming Conventions**

Function Name	Class	Description
~	Numerical Sample	STDEV or VAR. Ignores non-numbers. Divides sum by COUNT-1.
D~	Data Base Function	Operates on a contiguous range of cells as input
~P	Numerical Population	Ignores non-numbers. Divides sum by COUNT.
~A	Variant	Interprets TRUE as 1.0, FALSE as 0.0, any text as 0.0, numbers as their value and blanks are ignored. Uses the COUNTA function instead of COUNT. For STDEVA and VARA, the function divides by COUNTA-1.
~PA	Variant Population	Follows the ~A type interpretation and divides by COUNTA instead.

Note:

The ~A functions were added to allow spreadsheets generated under Lotus 1-2-3 to be read by Excel. The ~A functions represent design errors from a statistical viewpoint, since it violates the entire concept of “Levels of Measurement”. Do not use any ~A function in statistical analysis, since it combines all four (or five) levels into one meaningless summary number.

Variant is an object that can contain numbers, characters (strings, text), nulls, blanks or logic designators. Each cell in an Excel sheet represents one variant object. (See section 3)

The term “value” is anything in the cell. A blank cell has no “value”. COUNT gives a number for the count of “numbers” in the range. COUNTA gives a number for a count of “values” in the range.

The ~ and ~P functions, test each cell encountered, and will skip any blank cell, will skip any cell with a string/text variable and will skip any cell with a Boolean variable (TRUE or FALSE).

The NIST data sets for testing univariate functions only give values for the following statistics:

Sample Mean

Sample Standard Deviation

Sample Autocorrelation Coefficient

Excel does not have a Sample Autocorrelation Coefficient Function in the exact equation form given in the NIST definition of the equation. See Note L on this. The Excel CORREL function is different from the NIST equation, and should not be used for calculating lagged autocorrelation values. The values are different and will lead to failure of Excel's CORREL, as stated in McCullough and Wilson (1998). (See Note L on a possible new autocorrelation function in Excel.)

McCullough (2007) used the Excel CORREL function (a lag of one) on the NIST data, and compared the results to the results of a calculation corresponding to the correlation equation done in Mathematica. He just confirmed the inherent accuracy of the CORREL function.

Associated with descriptive statistics are graphics representing the data such as histograms, boxplots, stem and leaf plots, dot plots, pie charts, Pareto charts and others. Section 18 covers what Excel has in support of descriptive statistics.

## **UNIVARIATE DATA ANALYSIS – DATA ANALYSIS ROUTINES:**

### **FUNCTIONS/OUTPUTS PROVIDED**

The “Descriptive Statistics” routine in the Data Analysis add-in, provides univariate values. A typical output looks like this

**Figure 5-1: Descriptive Statistics Output Table**

<u>Column1</u>	
Mean	518.9587156
Standard Error	19.75639845
Median	522.5
Mode	671
Standard Deviation	291.6997275
Sample Variance	85088.73101
Kurtosis	-1.192560911
Skewness	-0.093331653
Range	995
Minimum	4
Maximum	999
Sum	113133
Count	218

The listed values are from the following functions. It only works with numeric data. An exception will be generated if the range contains non-numeric data.

**Table 5-2: Univariate Functions Called by “Descriptive Statistics”**

Row	Descriptive Statistics Line Title	Excel Functions Used
1	Mean	AVERAGE(-)
2	Standard Error	STDEV(-) / SQRT(COUNT(-))
3	Median	MEDIAN(-)
4	Mode	MODE(-)
5	Standard Deviation	STDEV(-)
6	Sample Variance	VAR(-)
7	Kurtosis	KURT(-)
8	Skewness	SKEW(-)
9	Range	MAX(-)-MIN(-)
10	Minimum	MIN(-)
11	Maximum	MAX(-)
12	Smallest	SMALL(-,k)
13	Largest	LARGE(-,k)
14	Sum	SUM(-)
15	Count	COUNT(-)
16	Confidence Level (95.0)	None. See notes below.

## **NOTES AND COMMENTS ON THE PROVIDED FUNCTIONS / OUTPUTS**

### **VARIANCE**

The value calculated is referred to as the “unbiased” value. This is the conventional variance value used for statistical calculations. It is considered as the “best” estimate of a population variance. If the data essentially is that of an entire population, then the VARP(-) function should be used instead of VAR(-). For data about an entire population then the descriptive statistics output would be wrong. Here you multiply the table output value by (n-1)/n for a population value. The NIST value is the conventional unbiased value.

### **STANDARD DEVIATION**

The value calculated (the square root of VAR) is referred to as a “biased” value, even though it comes from an “un-biased” variance estimate. The calculation done here is the conventional standard deviation value used throughout statistical practice. The NIST value is the conventional biased sample value.

There may be situations where an “unbiased” standard deviation is required<sup>1</sup> (See Note M, An Actual Problem Requiring Unbiased Standard Deviations).

---

<sup>1</sup> The words “biased” and “unbiased” have specific statistical definitions and meanings. In common language “biased” and “unbiased” have other meanings and connotations. When data or measurements are involved in legal situations, challenges to federal/state laws and regulations or end up as evidence in court cases, the common language meanings become the criteria. Consequently only the “unbiased” estimates may be admissible. or acceptable.

The conversion is:

Unbiased sigma = factor \* conventional biased sigma

Factor =  $(\Gamma(n/2) / [\sqrt{(n-1)/2} * \Gamma((n-1)/2)])$

Where  $\Gamma$  is the gamma function and  $\sqrt{\quad}$  is the square root function

The correction is small for n values above 20, and is one reason why it is not an important issue. Table 5-3 gives values of the factor for different sample sizes.

**Table 5-3: Standard Deviation Conversion Factor Values**

N	Factor
2	.797885
3	.886227
4	.921318
5	.939986
6	.951533
7	.959369
8	.965030
9	.969311
10	.972659
11	.975350
12	.977559
13	.979406
14	.980971
15	.982316
16	.983484
17	.984506
18	.985410
19	.986214
20	.986934

### **SKEWNESS AND KURTOSIS**

Excel calculates one version of the many ways these terms are defined and computed, by articles, papers and text books since 1895.

The Excel equation for skewness is defined (in help) as:

$$\frac{n}{(n-1)(n-2)} \sum \left( \frac{x_j - \bar{x}}{s} \right)^3$$

where n is the sample size, m is the mean (from AVERAGE) and s is the biased standard deviation (from STDEV)

The Excel equation for Kurtosis is defined (in help) as:

$$\left\{ \frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum \left( \frac{x_j - \bar{x}}{s} \right)^4 \right\} - \frac{3(n-1)^2}{(n-2)(n-3)}$$

where n is the sample size, m is the mean (from AVERAGE) and s is the biased standard deviation (from STDEV).

Skewness and kurtosis values involve concepts and ideas from Pearson (ca1905) and Fisher (ca 1925-1934). Let us start with the basic sums of the powers and mean values (m's) as defined by Fisher and corresponds to Pearson's usage (Heiser 2000).

$$\begin{aligned} m_1 &= \sum x_i / n \\ S_1 &= \sum x_i \\ S_2 &= \sum (x_i - m_1)^2 \\ S_3 &= \sum (x_i - m_1)^3 \\ S_4 &= \sum (x_i - m_1)^4 \\ m_2 &= S_2 / n \\ m_3 &= S_3 / n \\ m_4 &= S_4 / n \end{aligned}$$

Pearson originated the concepts and defined two basic moment ratios:

$$\begin{aligned} \beta_1 &= \mu_3^2 / \mu_2^3 \\ \beta_2 &= \mu_4 / \mu_2^2 \end{aligned}$$

In contemporary literature Pearson's measures are defined as:

$$\begin{aligned} \sqrt{b_1} &= m_3 / m_2^{(3/2)} \\ b_1 &= m_3^2 / m_2^3 \\ b_2 &= m_4 / m_2^2 \end{aligned}$$

The difference is due to Fisher who introduced the fact that sample measures are not population measures. Pearson took them to be the same. Fisher defined two other measures for samples as:

$$\begin{aligned} g_1 &= k_3 / k_2^{(3/2)} \\ g_2 &= k_4 / k_2^2 \end{aligned}$$

Where the k values are the semi-invariants:

$$\begin{aligned} k_1 &= m \\ k_2 &= [(n / (n - 1))] m_2 \\ k_3 &= [n^2 / (n-1)(n-2)] m_3 \\ k_4 &= \{n^2 / [(n-1)(n-2)(n-3)]\} \{(n + 1)m_4 - 3(n-1)m_2^2\} \end{aligned}$$

Pearson's and Fisher's measures are related by:

$$\sqrt{b_1} = (n-2) g_1 / \sqrt{[n(n-1)]}$$

Fisher's  $g$  measures are unbiased estimates of the population values. Pearson's are biased estimates. (Heiser 2000)

Excel calculates the  $g_1$  value. Use the above equation to convert to Pearson's value<sup>2</sup>.

There are three measures of kurtosis, Pearson's ( $b_2$ ), Fisher's ( $g_2$ ) and a third called "Excess Kurtosis". Excel calculates the  $g_2$  value. They are related by:

$$g_2 = \left\{ \frac{(n-1)}{[(n-2)(n-3)]} \right\} \left\{ (n+1) b_2 - 3(n-1) \right\}$$

$$b_2 = \left\{ \frac{g_2 (n-2)(n-3)}{[(n-1)(n+1)]} \right\} + 3(n-1)/(n+1)$$

The fourth moment of any distribution involves the constant 3. For a normal distribution  $b_2$  is 3. Pearson's  $b_2$  is centered about 3. Often 3 is subtracted from the computed  $b_2$  value to give a minus or positive kurtosis value. The term "excess kurtosis" usually means the  $b_2$  value minus 3. Fisher's  $g_2$  is unique and is centered around zero. Excess kurtosis is not applicable to Fisher's  $g_2$ .

### **CONFIDENCE LEVEL**

The confidence Level value here is different from the CONFIDENCE function value. The Data Analysis Confidence Level is a calculation of an interval about the sample mean using the biased sample standard deviation and a t value. The interval is one that is likely to contain the population mean at an alpha level of 0.05. The CONFIDENCE function calculates a half width interval about the sample mean using the population standard deviation and a z value.

These are two different values. This is not an error. It represents two different estimates from the sample of an interval defined in terms of population mean and standard deviation. Under repeated sampling from a fixed normal population, and both intervals as full widths, they will include the population mean in 95% of the time, taken as a long run of repeated sampling.

However in any given sample, both intervals may include the population mean, only one interval will include the population mean, or neither interval will contain the population mean. These intervals are not random values, but the process of obtaining a value for the interval is random, since it depends on sample values, which are random. The correct interpretation here is "if we repeat many times the process of taking a random sample of size  $n$  from this population, and compute a 95% confidence interval from each sample, then 95% of these confidence intervals will contain the true population mean" (Henderson and Meyer, 2001).

---

<sup>2</sup> Actually it is confusing, since Pearson gave several different measures of skewness, However the symbol ( $\sqrt{b_1}$ ) and equations, accurately states this symbolic ( $\sqrt{b_1}$ ) measure.

## **REPORTED PROBLEMS:**

The following is a summary table of problems with the univariate functions and routines, giving the application, the problem and a fix or workaround.

**Table 5-4: Excel Problems With Univariate Results**

<b>Application or Function</b>	<b>Problem</b>	<b>Source</b>	<b>Fix or Comments</b>
Univariate Statistics	Accuracy	McCullough 1999b, 2000	Fix A for 2000. Fix B for 2003
Standard Deviation (STDEV)	Incorrectly calculated.	Cryer 2001 Simon, CISE 27/99	Fix A for 2000. Fix B for 2003
Geometric Mean	GEOMEAN does not calculate correctly for large data sets.	Braden 2001a	Fix C
Quartiles	Cryer does not state what the problem is. Hunt says that differences between textbook percentiles and Excel have been noted.	Cryer 2001	Excel calculates correct quartiles and median values. No fix needed. See Note N
Ranking data	EXCEL does not treat tied observations correctly when ranking. Ties not handled properly	Cryer 2001, RSS 1996	The Data Analysis routine Rank and Percentile is confusing, <u>but does not have errors. There are many equal methods to deal with tied observations.</u>
Correlation Coefficient on Uni-variate data	Very low LRE values using the NIST StDR tests	McCullough 2000	Fix D

## **RESULTS OF ACCURACY TESTS:**

### **DIFFERENCES BETWEEN EXCEL VERSIONS**

Excel versions 97 and 2000 represent a common set of algorithms used in all the functions. The results from Excel 2000 testing are also valid for Excel 97.

Excel versions 2003 and 2007 represent a different common set of algorithms used in all the functions. The results from Excel 2003 testing are also valid for Excel 2007.

Tables of the results from 2000 testing will be different from 2003 testing. Not only in values, but in the test data. Excel 2000 could not be backloaded, so that the tests done ended up as being different.

## **TEST GROUP 1:**

### **TESTS CONDUCTED**

Table 5-5 gives the LRE results from Excel reported by McCullough (1997, 1998 and 2000) on the NIST univariate data sets. This was on the 97-2000 version. Also included (for comparison) are LRE values for some commercial software packages reported by McCullough (1997, 1998 and 2000) and Creighton and Ding (2002) on the same sets. Only standard deviations are reported.

The values reported by Altman (2002) on JMP 4.0.3 were contested by the SAS Institute (Sall 2002). They claimed that they got “very different accuracy scores than Altman”. However Sall (2002) was not very clear on what actual values reported by Altman were to their point of view inaccurate. Creighton, Ding, Sall and Gotwalt (2002) expanded on the issue of Altman’s review, based on the tests conducted by Creighton and Ding (2002) on JMP 4.0.5. Their published values are included in the tables below, as a comparison of Excel’s outputs with another commercial statistical software package.

Also included are my results using the same data sets in Excel 2000, using the fixes described in this article.

**Table 5-5: Univariate Analysis Results on Excel 2000 and 2003 with StRD data sets**

<b>Sequence</b>	<b>Source</b>	<b>Dataset</b>	<b>Category</b>	<b>Difficulty</b>	<b>Size</b>	<b>Number of Significant Figures</b>	<b>LIC Value</b>
1	NIST	PiDigits	Univariate	1	5000	1	0.1991
2	NIST	Lottery	Univariate	1	218	3	0.2502
3	NIST	Lew	Univariate	1	200	3	0.1940
4	NIST	Maveo	Univariate	1	50	5	3.6689
5	NIST	Michelso	Univariate	1	100	5	3.5792
6	NIST	NumAcc1	Univariate	1	3	8	7.0000
7	NIST	NumAcc2	Univariate	2	1001	2	1.0792
8	NIST	NumAcc3	Univariate	2	1001	8	7.0000
9	NIST	NumAcc4	Univariate	3	1001	9	8.0000

**Table 5-5: Univariate Analysis Results on Excel 2000 and Excel 2003 with StRD data sets (continued)**

Sequence	Excel 2000, McCullough 2000			Stata, McCullough 2000			JMP, Creighton & Ding 2002		
	Mean	Standard Deviation	First Order Correlation Coefficient	Mean	Standard Deviation	First Order Correlation Coefficient	Mean	Standard Deviation	First Order Correlation Coefficient
1	15.0	15.0	4.0	15.0	15.0	14.9	15.0	15.0	13.0
2	15.0	15.0	2.1	15.0	15.0	15.0	15.0	15.0	15.0
3	15.0	15.0	2.6	15.0	15.0	14.8	15.0	15.0	15.0
4	15.0	9.4	1.8	15.0	13.1	13.7	15.0	13.1	13.8
5	15.0	8.3	3.6	15.0	13.8	13.4	15.0	13.8	13.4
6	15.0	15.0	0.0	15.0	15.0	15.0	15.0	15.0	15.0
7	14.0	11.6	3.3	15.0	15.0	15.0	14.0	14.6	13.7
8	15.0	1.1	3.3	15.0	9.5	11.9	15.0	9.5	11.2
9	14.0	0.0	3.3	15.0	8.3	10.7	15.0	8.3	9.0

**Table 5-5: Univariate Analysis Results on Excel 2000 and 2003 with StRD data sets (Continued)**

Sequence	Excel 2000, Centered		Excel 2003		Xnumbers-46	
	Mean	Standard Deviation	Mean	Standard Deviation	Mean	Standard Deviation
1	15.0	15.0	15.0	15.0	15.0	15.0
2	15.0	15.0	15.0	15.0	15.0	15.0
3	15.0	15.0	15.0	15.0	15.0	15.0
4	15.0	9.4	15.0	13.1	15.0	15.0
5	15.0	8.3	15.0	13.8	15.0	15.0
6	15.0	15.0	15.0	15.0	15.0	15.0
7	14.0	15.0	14.0	11.6	15.0	15.0
8	15.0	12.8	15.0	9.5	15.0	15.0
9	14.0	12.2	14.0	8.3	15.0	15.0

The outstanding accuracy of the xnumbers add-in (see note XA ) is due to the conversion of double precision numbers to character strings. This allows an unlimited number of digits to represent the number, but requires special Excel functions to achieve the high accuracies. In this particular case, 30 decimal digits (as the floating point mantissa)

results in computed values in all cases (including the tables below) having LRE values of 15.0 or higher.

## **TEST GROUP 2:**

### **TESTS CONDUCTED**

NIST only provides precision values for the mean, standard deviation and first order correlation coefficient. To test the other univariate functions then, additional test data has to be selected, and reference values generated to compare with the results from Excel. The Michelson data set was selected, and special vba functions were written to establish reference values. This is a sort of pulling-yourself-up-by-your-own-shoelaces approach. I had no access to any method that did not rely on the IEEE-754 floating-point limitation. Consequently the reader must recognize that actual Excel LRE values may be less than the values reported in the following tables.

The tests were on variations of the Michelson Data set, sequence 5 of the NIST data set. The variations were to change the order (as given, sorted ascending and sorted descending) and to compare centered data to as-is data.

### **TEST RESULTS**

Table 5-6, shows that arranging the data in some order has an effect on the Excel function accuracies. It also shows the improvement in accuracy that results when the data is first centered. LRE values above 15 are shown only to indicate subtle changes in the binary mantissa sequences. All LRE values of 15 and above should be considered fully accurate.

**Table 5-6: Excel 2000 Data Analysis (add-in) Descriptive Statistics on the Michelson Data Set, LRE Values**

Property	As Given		Sorted Ascending		Sorted Descending	
	Normal	Centered	Normal	Centered	Normal	Centered
Mean	15.72	15.72	16.00	15.72	15.42	15.72
Standard Error*	8.28	13.85	9.11	13.85	9.14	13.85
Median	16.00	10.85	16.00	10.85	16.00	10.85
Mode	16.00	12.61	16.00	12.61	3.63	12.61
Standard Deviation*	8.28	13.85	9.11	13.85	9.14	13.85
Sample Variance*	7.98	13.56	8.81	13.56	8.84	13.56
Kurtosis*	12.37	12.20	12.25	12.23	12.06	12.23
Skewness*	9.92	10.67	11.66	10.67	9.63	10.67
Range	13.60	13.60	13.60	13.60	13.60	13.60
Minimum	16.00	13.17	16.00	13.17	16.00	13.17
Maximum	16.00	13.69	16.00	13.69	16.00	13.69
Sum	16.00	11.94	16.00	11.94	16.00	11.94
Count	16.00	16.00	16.00	16.00	16.00	16.00

Doing the same analysis in Excel 2003 resulted in the values in table 5-7.

**Table 5-7: LRE Values, Excel 2003 Data Analysis Descriptive Statistics on the Michelso Data Set**

<b>Property</b>	<b>As Given</b>	<b>Sorted</b>	<b>Sorted</b>
Mean	15.72	16.00	15.42
Standard Error*	13.83	13.85	13.85
Median	16.00	16.00	16.00
Mode	16.00	13.65	16.00
Standard Deviation*	13.84	13.85	13.85
Sample Variance*	13.55	13.56	13.56
Kurtosis*	12.37	12.25	12.06
Skewness*	9.92	11.66	9.63
Range	13.60	13.60	13.60
Minimum	16.00	16.00	16.00
Maximum	16.00	16.00	16.00
Sum	16.00	16.00	16.00
Count	16.00	16.00	16.00

The effects of different sorting arrangements have been reduced. The standard error, standard deviation and sample variance values no longer are sort dependent. However the Mode, Mean, Kurtosis and Skewness values are still sort order dependent.

Microsoft in KBA-214282 noted the problem of getting different answers depending on the sequence of values.

**TEST GROUP 3:**  
**TESTS CONDUCTED**

The question then comes up, when do the standard univariate functions loose their accuracies?

Tables -A, -B and -C below represent the results of the standard univariate functions on an offset data series (Section 4 describes these offset data sets.). The GEOMEAN and HARMEAN functions were left out.

Table A: 1001 uniform distributed values of the offset type type (standard deviation of 0.100018) were created as a baseline. For the 1001 set, the lower bound was 0.026795 and the upper bound was 0.373205.

Table B: 1001 values of the offset data type (one 0.2 value and alternating 0.1 and 0.3 values) were created as a baseline

Table C: 1001 random normal deviates of the offset type (mean of 0.2, sigma = 0.1) were created outside of Excel as a baseline

The distributions were adjusted so that each had a mean of 0.2 and a standard deviation of 0.1 to correspond to the NumAcc style of data.

An additive was added to the random values, increasing the number of significant digits in columns 3 to 10 along the same lines as the NumAcc3 and NumAcc4 data sets change. Column 2 represents values obtained from the functions on the basic variations without

any additive. Columns 3 through 10 represent function outputs on the column 1 data set with the additive given in the first row added to each point. Row 3 is the L10COV value for the set.

Table set 5-8 represents values from Excel 2000 without centering.

Table set 5-9 represents values from Excel 2000 with the data pre-centered.

Table set 5-10 represents values from Excel 2003 without centering.

**TEST RESULTS**

**Table 5-8-A: Excel 2000 LRE Values. Deviations Uniformly Distributed**

<b>Additive</b>	<b>0</b>	<b>1</b>	<b>10</b>	<b>100</b>	<b>1000</b>	<b>10000</b>	<b>100000</b>	<b>1000000</b>	<b>10000000</b>
LIC Value	0.30	1.08	2.01	3.00	4.00	5.00	6.00	7.00	9.00
AVERAGE	15.00	15.00	15.00	14.95	15.00	15.00	15.00	15.00	15.00
AVEDEV	15.00	15.80	15.00	14.80	14.75	13.19	11.91	11.58	9.60
DEVSQ	15.00	15.05	15.00	14.80	14.57	13.30	11.75	12.55	9.47
KURT	14.62	15.43	14.56	14.73	15.73	13.04	11.56	11.95	8.81
MAX	15.00	15.00	15.00	15.00	15.00	15.00	15.00	15.00	15.00
MEDIAN	15.00	15.00	15.00	15.00	15.00	15.00	15.00	15.00	15.00
MIN	15.00	15.00	15.00	15.00	15.00	15.00	15.00	15.00	15.00
SKEW	15.00	14.76	13.07	11.48	11.69	10.66	10.10	8.25	7.05
STDEV	15.00	13.11	12.75	9.01	8.47	6.39	3.74	1.50	0.00
STDEVP	14.95	13.15	13.08	8.99	8.53	6.81	3.74	1.47	0.00
VAR	14.81	12.81	12.44	8.71	8.17	6.09	3.44	1.20	0.00
VARP	14.68	12.84	12.78	8.69	8.23	6.51	3.44	1.17	0.00

**Table 5-8-B: Excel 2000 LRE Values. Deviations NumAcc Distributed**

<b>Additive</b>	<b>0</b>	<b>1</b>	<b>10</b>	<b>100</b>	<b>1000</b>	<b>10000</b>	<b>100000</b>	<b>1000000</b>	<b>10000000</b>
LIC Value	0.30	1.08	2.01	3.00	4.00	5.00	6.00	7.00	8.00
AVERAGE	14.74	14.03	14.13	13.81	15.00	14.05	13.82	15.24	14.01
AVEDEV	13.86	14.60	13.87	13.52	12.48	11.34	12.14	9.45	8.18
DEVSQ	13.91	14.29	13.97	13.80	12.17	10.96	10.54	9.16	7.95
KURT	14.31	14.51	14.88	13.79	14.65	14.88	14.65	14.38	11.71
MAX	15.00	15.00	15.00	15.00	15.00	15.00	15.00	15.00	15.00
MEDIAN	15.00	15.00	15.00	15.00	15.00	15.00	15.00	15.00	15.00
MIN	15.00	15.00	15.00	15.00	15.00	15.00	15.00	15.00	15.00
SKEW	13.99	12.47	11.64	10.33	10.59	8.58	7.34	7.72	5.53
STDEV	13.54	11.57	9.76	7.64	7.22	5.82	2.10	1.14	0.00
STDEVP	13.54	11.57	9.76	7.64	7.21	5.81	2.10	1.14	0.00
VAR	13.23	11.27	9.46	7.34	6.92	5.52	1.80	0.82	0.00
VARP	13.23	11.27	9.46	7.34	6.91	5.51	1.80	0.82	0.00

**Table 5-8-C: Excel 2000 LRE Values. Deviations N(0.2,0.1) Distributed.**

Additive	0	1	10	100	1000	10000	100000	1000000	10000000
LIC Value	0.30	1.08	2.01	3.00	4.00	5.00	6.00	7.00	8.00
AVERAGE	15.00	15.00	15.00	15.00	15.00	14.96	15.00	15.00	15.00
AVEDEV	15.00	15.00	14.61	13.74	13.00	11.40	10.73	10.11	8.71
DEVSO	15.00	15.00	15.00	14.97	13.88	12.34	11.54	10.92	9.69
KURT	15.00	13.73	13.05	11.89	11.10	9.43	8.78	8.26	6.83
MAX	15.00	15.00	15.00	15.00	15.00	15.00	15.00	15.00	15.00
MEDIAN	15.00	15.00	15.00	15.00	15.00	15.00	15.00	15.00	15.00
MIN	15.00	15.00	15.00	15.00	15.00	15.00	15.00	15.00	15.00
SKEW	15.00	13.09	12.14	10.97	10.18	8.58	7.89	7.33	5.86
STDEV	15.00	13.23	11.84	9.30	8.25	5.01	3.04	1.44	0.00
STDEVP	15.00	13.26	11.97	9.30	8.22	5.01	3.03	1.45	0.00
VAR	15.00	12.93	11.54	9.00	7.95	4.71	2.73	1.14	0.00
VARP	15.00	12.96	11.67	9.00	7.92	4.71	2.73	1.14	0.00

The standard deviation function STDEV is known to be bad in all versions before Excel 2003. Based on the above tables, an accurate variance or standard deviation cannot be obtained if the L10COV value is larger than 5. The type of distribution also has an effect.

Table set 5-9 represents the same functions on the same data set, except the data sets were centered about their mean. It is evident that centering improves the accuracy. Centered data has no L10COV values.

**Table 5-9-A: Excel 2000 LRE Values. Deviations Uniformly Distributed, Data Centered Before Analysis**

Additive	0	1	10	100	1000	10000	100000	1000000	10000000
AVEDEV	15.00	15.00	15.00	14.03	14.08	13.19	11.92	11.68	9.55
DEVSO	15.00	15.00	14.91	14.71	14.71	13.29	11.75	12.55	9.47
KURT	14.62	15.00	14.50	15.00	14.53	13.07	11.56	11.95	8.81
MAX	15.00	15.02	13.87	12.18	12.49	11.16	10.25	9.06	7.36
MEDIAN	15.00								
MIN	15.00	15.00	13.76	12.18	12.48	11.44	10.55	8.94	7.37
SKEW	15.00	15.00	14.98	15.00	14.22	13.72	11.19	11.60	9.45
STDEV	15.00	15.00	15.00	14.31	14.27	13.68	12.05	12.84	9.77
STDEVP	15.00	15.00	15.00	15.00	14.95	13.60	12.05	12.86	9.77
VAR	15.00	15.00	15.00	14.81	14.56	13.30	11.75	12.55	9.47
VARP	15.00	15.00	15.00	14.98	14.65	13.30	11.75	12.56	9.47

**Table 5-9-C: Excel 2000 LRE Values. Deviations Normally Distributed, Data Centered Before Analysis**

<b>Additive</b>	<b>0</b>	<b>1</b>	<b>10</b>	<b>100</b>	<b>1000</b>	<b>10000</b>	<b>100000</b>	<b>1000000</b>	<b>10000000</b>
AVEDEV	15.00	15.00	15.00	15.00	15.00	14.80	13.08	12.27	11.09
DEVSO	15.00	15.00	14.97	14.41	13.88	12.34	11.54	10.92	9.69
KURT	14.57	14.57	14.15	13.62	13.37	11.24	11.65	10.31	8.48
MAX	15.00	15.00	14.45	13.32	12.61	11.03	10.36	9.63	8.21
MEDIAN	15.00	15.00	14.60	13.38	12.51	10.99	10.28	9.70	8.36
MIN	15.00	15.00	14.54	13.46	12.69	10.97	10.36	9.72	8.31
SKEW	14.23	14.22	14.23	14.51	13.44	12.11	11.65	10.67	8.80
STDEV	15.00	15.00	15.00	15.00	15.00	13.65	12.84	12.22	10.99
STDEVP	15.00	15.00	15.00	15.00	15.00	13.66	12.84	12.22	10.99
VAR	15.00	15.00	15.00	15.00	15.00	14.35	13.54	12.92	11.69
VARP	15.00	15.00	15.00	15.00	15.00	14.35	13.54	12.92	11.69

In tables 5-9-A and 5-8-C, the loss in accuracy of the MAX, MIN and MEDIAN values represent the fact that the difference between the mean and the value, as the additive increases, loses accurate digits. This occurs from the finite limit of 15 decimal digits. When the additive is 8 decimal digits, then only the higher 6-7 digits of the difference are accurate.

Centering the data first markedly improves the accuracy.

Applying the same tests to Excel 2003 gives the following results

**Table 5-10-A-: Excel 2003 LRE Values. Deviations Uniformly Distributed**

<b>Additive</b>	<b>0</b>	<b>1</b>	<b>10</b>	<b>100</b>	<b>1000</b>	<b>10000</b>	<b>100000</b>	<b>1000000</b>	<b>10000000</b>
LIC Value	0.30	1.08	2.01	3.00	4.00	5.00	6.00	7.00	9.00
AVERAGE	15.00	15.00	15.00	14.95	15.94	16.00	16.00	15.63	16.00
AVEDEV	15.00	15.00	15.00	14.80	14.75	13.19	11.91	11.58	9.60
DEVSO	15.00	15.00	15.00	14.80	14.57	13.30	11.75	12.55	9.47
KURT	14.62	15.00	14.56	14.73	15.73	13.04	11.56	11.95	8.81
MAX	15.00	15.00	15.00	15.00	15.00	15.00	15.00	15.00	15.00
MEDIAN	15.00	15.00	15.00	15.00	15.00	15.00	15.00	15.00	15.00
MIN	15.00	15.00	15.00	15.00	15.00	15.00	15.00	15.00	15.00
SKEW	15.00	14.76	13.07	11.48	11.69	10.66	10.10	8.25	7.05
STDEV	15.00	13.11	16.00	14.31	14.27	13.68	12.05	12.84	9.77
STDEVP	15.00	13.11	16.00	15.26	14.95	13.60	12.05	12.86	9.77
VAR	14.81	12.81	15.46	14.81	14.56	13.30	11.75	12.55	9.47
VARP	14.81	12.81	16.00	14.92	14.65	13.30	11.75	12.56	9.47

**Table 5-10-B: Excel 2003 LRE Values. Deviations NumAcc Distributed**

<b>Additive</b>	<b>0</b>	<b>1</b>	<b>10</b>	<b>100</b>	<b>1000</b>	<b>10000</b>	<b>100000</b>	<b>1000000</b>	<b>10000000</b>
LIC Value	0.30	1.08	2.01	3.00	4.00	5.00	6.00	7.00	8.00
AVERAGE	14.74	14.03	14.13	13.81	15.00	14.05	13.82	15.24	14.01
AVEDEV	13.86	14.60	13.87	13.52	12.48	11.34	12.14	9.45	8.18
DEVSO	13.91	14.29	13.97	13.80	12.17	10.96	10.54	9.16	7.95
KURT	14.31	14.51	14.88	13.79	14.65	14.88	14.65	14.38	11.71
MAX	15.00	15.00	15.00	15.00	15.00	15.00	15.00	15.00	15.00
MEDIAN	15.00	15.00	15.00	15.00	15.00	15.00	15.00	15.00	15.00
MIN	15.00	15.00	15.00	15.00	15.00	15.00	15.00	15.00	15.00
SKEW	13.99	12.47	11.64	10.33	10.59	8.58	7.34	7.72	5.53
STDEV	13.54	11.57	14.28	14.10	12.47	11.26	10.84	9.46	8.25
STDEVP	13.54	11.57	14.27	14.10	12.47	11.26	10.84	9.46	8.25
VAR	13.23	11.27	13.97	13.80	12.17	10.96	10.54	9.16	7.95
VARP	13.23	11.27	13.97	13.80	12.17	10.96	10.54	9.16	7.95

**Table 5-10-C: Excel 2003 LRE Values. Deviations N(0.2,0.1) Distributed.**

<b>Additive</b>	<b>0</b>	<b>1</b>	<b>10</b>	<b>100</b>	<b>1000</b>	<b>10000</b>	<b>100000</b>	<b>1000000</b>	<b>10000000</b>
LIC Value	0.30	1.08	2.01	3.00	4.00	5.00	6.00	7.00	8.00
AVERAGE	15.00	15.00	15.00	15.00	15.00	14.96	15.36	15.63	15.25
AVEDEV	15.00	15.00	14.61	13.74	13.00	11.40	10.73	10.11	8.71
DEVSO	15.00	15.00	15.00	14.97	13.88	12.34	11.54	10.92	9.69
KURT	15.00	13.73	13.05	11.89	11.10	9.43	8.78	8.26	6.83
MAX	15.00	15.00	15.00	15.00	15.00	15.00	15.00	15.00	15.00
MEDIAN	15.00	15.00	15.00	15.00	15.00	15.00	15.00	15.00	15.00
MIN	15.00	15.00	15.00	15.00	15.00	15.00	15.00	15.00	15.00
SKEW	15.00	13.09	12.14	10.97	10.18	8.58	7.89	7.33	5.86
STDEV	15.00	13.23	14.38	14.44	14.48	12.65	11.84	11.22	9.99
STDEVP	15.00	13.23	14.38	14.44	14.68	12.66	11.84	11.22	9.99
VAR	15.00	12.93	14.08	14.14	14.48	12.35	11.54	10.92	9.69
VARP	15.00	12.93	14.08	14.14	15.46	12.35	11.54	10.92	9.69

Excel 2003 gives more accurate univariate results than Excel 2000. KURT and SKEW remain the most inaccurate of the univariate set. No changes were made to either of these functions. Both present difficulties in obtaining accurate values. They should have been included in the 1-2 pass modifications, since table 5-10-C shows that pre-centering can improve the accuracy of KURT and SKEW.

Table 5-11 compares some of the results on computation of the standard deviation. The data set for the Excel 2000, uncentered and centered was the same. The Excel 2003 runs included two different random number data sets from the MWC256 (Note AC).

Welford's algorithm is a one-pass alternate algorithm that corrects for addition rounding errors. It is discussed in Note P.

**Table 5-11: Standard Deviation LRE Values, N(0.2,0.1) Distributed Data**

<b>Additive</b>	<b>10</b>	<b>100</b>	<b>1000</b>	<b>10000</b>	<b>100000</b>	<b>1000000</b>	<b>10000000</b>
LIC Value	2.00	3.00	4.00	5.00	6.00	7.00	8.00
Excel 2000 STDEV	11.8	9.3	8.3	5.0	3.0	1.4	0
Excel 2000, Centered, STDEV	15.0	15	15.0	13.7	12.8	12.2	11.0
Excel 2003, STDEV, Set 1	14.4	14.4	14.5	12.7	11.8	11.2	10.0
Excel 2003, STDEV, Set 2	14.6	13.8	13.2	12.0	11.3	9.9	9.9
Welford's Algorithm, Set 2	13.9	13.2	12.0	10.5	10.5	9.0	8.6

Excel 2003 is an improvement over Excel 2000. The sum of the LOG10(constant added) value and the LRE value runs from 17 to 19 for the best accuracy method, runs from 16 to 18 for Excel 2003, and is about 16 for Welford's algorithm.

Table 5-10-C shows that Microsoft overlooked the obvious again. By using a first pass pre-centering was good, but the second pass change to just squaring the deviations was poor. By retaining the calculator equation, there is a built in automatic correction for addition rounding errors and the fact that the sum of the differences are not zero. The improvement in accuracy for long lists is one decimal digit, if the old calculator equation is used in the second pass.

There are however situations where the data cannot be pre-centered. The GEOMEAN and HARMEAN functions cannot be applied to centered data, and have remained the same. Microsoft never recognized Braden's fix.

#### **TEST GROUP 4:**

These were tests run on the Analysis Tool Pac Descriptive Statistics Routine, This routine just calls Excel univariate functions and outputs the values in a table. Table 5-2 above lists these functions and shows what the table looks like.

For this test, an offset data series was made up based on the NIST NumAcc data base. The magnitude was changed by orders of 10, giving LIC values from 0.3 to 16. Each set had 1001 data items, corresponding to the NumAcc series. The test series were run in Excel 2003 and in Excel 2007. There was no difference between the Excel 2003 and Excel 2007 results. The table from Excel 2007 is shown below.

**Table 5-12A: Descriptive Statistics Output Accuracy LRE Values, Excel 2007**

LIC Value	0.30	2.01	3.00	4.00	5.00	6.00	7.00	8.00	9.00	10.00	11.00
Mean	14.74	14.13	13.81	15.10	14.05	13.82	15.24	14.01	13.83	15.45	13.97
Standard Error	13.53	14.27	14.10	12.47	11.26	10.84	9.46	8.25	7.82	6.45	5.29
Median	16.00	16.00	16.00	16.00	16.00	16.00	16.00	16.00	16.00	16.00	16.00
Mode	16.00	16.00	16.00	16.00	16.00	16.00	16.00	16.00	16.00	16.00	16.00
Standard Deviation	13.54	14.28	14.10	12.47	11.26	10.84	9.46	8.25	7.82	6.45	5.29
Sample Variance	13.23	13.97	13.80	12.17	10.96	10.54	9.16	7.95	7.52	6.15	4.99
Kurtosis	14.31	14.88	13.79	14.65	14.88	14.65	14.38	11.71	9.35	10.46	5.65
Skewness	13.99	11.64	10.33	10.59	8.58	7.34	7.72	5.53	4.35	4.90	2.50
Range	15.86	14.28	13.85	12.47	11.26	10.84	9.46	8.25	7.83	6.45	5.24
Minimum	16.00	16.00	16.00	16.00	16.00	16.00	16.00	16.00	16.00	16.00	16.00
Maximum	16.00	16.00	16.00	16.00	16.00	16.00	16.00	16.00	16.00	16.00	16.00
Welford's	15.16	14.23	13.86	12.46	11.26	10.84	9.45	8.25	7.83	6.44	5.24

**Table 5-12B: Descriptive Statistics Output Accuracy LRE Values, Excel 2007**

LIC Value	12.00	13.00	14.00	15.00	16.00
Mean	14.00	16.00	13.98	14.73	15.90
Standard Error	-1.00	-1.00	-1.00	0.03	0.94
Median	16.00	16.00	16.00	16.00	16.00
Mode	16.00	16.00	16.00	16.00	15.00
Standard Deviation	-1.00	-1.00	-1.00	0.03	0.94
Sample Variance	-1.00	-1.00	-1.00	-1.00	-1.00
Kurtosis	3.36	6.13	0.30	0.43	0.30
Skewness	1.35	2.73	-0.15	-0.11	-0.15
Range	4.82	5.24	2.23	1.81	0.43
Minimum	16.00	16.00	16.00	16.00	16.00
Maximum	16.00	16.00	16.00	16.00	16.00
Welfords	4.82	3.44	2.23	1.73	0.94

The Excel functions STDEV, VAR, SKEW and KURT show a loss in accuracy here as the LIC value of the data set increases. The accuracy of the Range and Standard Error values also decrease.

A plot of the mean and standard deviation LRE values versus LIC values is in section 4, figure 4-1.

In section 4, it was pointed out that VAR and STDEV give wrong values when the LIC value exceeds about 10. This is strictly a problem with the DEVSQ algorithm. Other data sets show inconsistent behavior with the VAR and STDEV function when the magnitude of the whole number (left of the decimal point) is from 0 to 8, and a total loss of accuracy when the magnitude is greater than 10. Table 5-5 shows this. This is the basic Descriptive Statistics run on the expanded NumAcc data set.

The table also shows the slight improvement with Welford's exact algorithm with Kahan's modification to control summation errors. The real improvement with Welford's, is the consistency and evenness of the loss in accuracy as the LIC values increase

## **COMMENTS, FIXES, WORKAROUNDS AND RECOMMENDATIONS**

### **PROBLEMS WITH VARIANCE CLASS FUNCTIONS, EXCEL 97 AND 2000**

#### ***OVERVIEW AND COMMENTS***

The most frequent problem is about the accuracy of the VAR and STDEV functions. They are included internally in many other Excel functions and routines, consequently affecting the accuracy of these other functions and routines.

#### ***FIXES, WORKAROUNDS AND RECOMMENDATIONS:***

In Excel 97 and 2000, center the data first. The old calculator equation on the centered data will then give correct values.

### **PROBLEMS WITH VARIANCE CLASS FUNCTIONS, EXCEL 2003 AND 2007**

#### ***OVERVIEW AND COMMENTS***

Microsoft changed the VAR and STDEV functions for Excel 2003. They refer to the change as the "two-pass method", in which the data is first centered (pass 1) and then the sum of the squares of the deviations from the average are calculated (pass 2). This change improved the accuracy of many other functions that depended on a variance value of a set of numbers. All previous versions of Excel used the old one-pass calculator algorithm, shown in the Help section of the pre-2003 versions.

There are several KBA's (826112, 826349, 828888) that describe the change. For the VAR function it is:

$$\text{VAR} = \text{DEVSQ}(\text{range}) / (\text{COUNT}(\text{range}) - 1)$$

DEVSQ does the two passes. STDEV just takes the square root of the VAR value.

Testing has shown that DEVSQ and COUNT correctly skip blanks and any non-numeric data in any cell in the range.

However the new functions do not have the accuracy and consistency of values obtained by first centering the data and then applying the **old one-pass calculator function**. This is due to the fact that the sum of the centered values (as a sum of {IF} values) does not exactly come to zero (i.e. the true {IR} sum value is zero). This is shown in figure 4-1 as the LIC value of the data set increases. The old one-pass algorithm on centered data corrects for this. Welford's algorithm (note P) seems to be able to correct for this and will generally obtain fully accurate standard deviation values. See also Note O on this problem.

#### ***FIXES, WORKAROUNDS AND RECOMMENDATIONS:***

If an exact value is needed, first calculate an average, set up a column of differences from this average value and then apply the VAR or STDEV function to the column of differences.

## **PROBLEMS WITH THE GEOMETRIC MEAN EXCEL 2000 AND 2003**

### ***OVERVIEW AND COMMENTS***

The Excel geometric mean rapidly reaches the limit of floating point number size before the calculation on the range is finished. This occurs in all versions of Excel.

### ***FIXES, WORKAROUNDS AND RECOMMENDATIONS:***

Braden's algorithm (below) corrects the problem.

=EXP(AVERAGE(LN(rng))) where rng is the range of the data.

## **PROBLEMS WITH RANK AND PERCENTILE**

### ***OVERVIEW AND COMMENTS***

I found no error in RANK on the Michelson and other data sets. The Data Analysis routine Rank and Percentile is confusing, but does not have errors. The problem is with tied observations:

1. Problem 1: Given n ties at some point, do all the ties have the same rank? If the answer is yes, then problem 2 has to be dealt with. If the answer is no, then how are they distinguished from a univariate viewpoint?
2. Problem 2: Given m data values, does the ranking go from 1 to m? If the answer is yes, then how are m-n +1 values to be ranked, when they are obviously less than m. This is the incompatibility that ties create. If the answer is no, then the n ties are reduced in the analysis to a single value.

There are many equal methods to deal with tied observations. Dealing with tied observations is a problem. There is no standard, accepted method of interpretation as related to rank. All the alternate methods say that tied observations are not equal. See Note N.

### ***FIXES, WORKAROUNDS AND RECOMMENDATIONS:***

None.

## **PROBLEMS WITH AUTOCORRELATION**

### ***OVERVIEW AND COMMENTS***

The StRD reference sheets on the univariate data sets state that the univariate correlation being used is "Autocorrelation Coefficient (lag 1) r(1):" The StRD formula for the autocorrelation is:

$$\text{Tau} = \frac{\sum_2^n \{(Y_i - Y_m) \times (Y_{i-1} - Y_m)\}}{\{\sum_1^n (Y_i - Y_m)^2\}}$$

Where:  $Y_m$  is the average of the 1 to n data set.

Excel does not have this function. Therefore this measure cannot be tested for in Excel.

McCullough (1997, 1998 and 2000) in their tests used the Excel correlation function for this value. This is the CORREL function

$$\text{Correl}(X, Y) = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\sum (x - \bar{x})^2 \sum (y - \bar{y})^2}}$$

Although it appears similar to the NIST equation, however the values come out different from the NIST values, just enough to not pass the NIST tests. McCullough (2000) concluded that Excel's CORREL function gives wrong values when it is used to do a lag-1 autocorrelation on univariate data.

***FIXES, WORKAROUNDS AND RECOMMENDATIONS:***

A vba function can be written to do exactly what the NIST autocorrelation does. See Note L.

**CONCLUSIONS**

Excel 2003 and Excel 2007 are preferred over Excel 2000 on Univariate data analysis. The results are accurate to the limits given in tables 5-12. For users of Excel 2000, pre-centering the data first is recommended in all cases (except GEOMEAN and HARMEAN) to ensure accuracy.

Where higher accuracies are required, one can use the xnumber process, but the sacrifice is longer computing times.