

VIII. COVARIANCE AND CORRELATION.....	1
OVERVIEW OF THE PROBLEM	1
EXCEL FUNCTIONS:	2
EXCEL DATA ANALYSIS TOOLPAC ROUTINES.....	2
CORRELATION	2
COVARIANCE	3
EXCEL 97 AND 2000	3
REPORTED PROBLEMS.....	3
RESULTS OF TESTS	3
CONCLUSIONS.....	7
RECOMMENDATIONS	7
EXCEL 2003 AND 2007	8
CHANGES FROM EXCEL 2000 IN EXCEL 2003 AND 2007	8
RESULTS OF TESTS ON EXCEL 2003.....	8
MISSING DATA	9
RECOMMENDATIONS	9

VIII. COVARIANCE AND CORRELATION

OVERVIEW OF THE PROBLEM

The whole area of correlations and correlation measures is poorly covered in textbooks. Out of a survey of ten textbooks, I found over ten different equations. There seems to be no attempt to arrive at a commonality in how this area is presented to students. Out of these nine textbooks, there were six different definitions of correlation. Some textbooks listed in their index only under coefficients and some only under correlations.

I took these textbook equations and wrote vba functions corresponding to the equations. The Norris StRD data set was used as the input data set. Out of all the different equations, there were only 3 different values:

#1=0.999996872936967

#2=0.999993745883712

#3=0.972219182022051

The following table shows what I found.

Table 8-1: Correlation Measure Values, Norris Data Set

Source	Correlation	Correlation Coefficient	Coefficient of Correlation	Coefficient of Determination	Square of the Correlation	Pearson Product Moment Correlation
M&M	#1				#2	
Larson		#1		#2		
Levine		#1	#1	#2		
Pelosi		#1				
Neter			#1	#2		#1
Bollen		#1				
Pearl		#1				
Dretzke						#1
Mendenhall		#1				
Ostle			#1	#2		

The correct value based on the StRD reference is #2 for the coefficient of determination (R^2 or r^2 depending on the textbook). #1 is the square root of #2.

EXCEL FUNCTIONS:

CORREL – Returns the correlation coefficient between two separate ranges of paired data. It is the r-value (#1 value). Skips pairs that involve cells with missing data.

COVAR – Returns the covariance between two separate ranges of paired data.

PEARSON – Returns the Pearson product moment correlation between two separate ranges of paired data. For Excel 2003 and 2007, the returned value is identical to CORREL and the Help equations are identical. Skips pairs that involve cells with missing data.

RSQ – Returns the squared Pearson product moment correlation between two separate ranges of paired data. This is the #2 value.

MCORREL – Returns a correlation matrix. Can only be used as a macro in a macro sheet.

MCOVAR – Returns a covariance matrix. Can only be used as a macro, in a macro sheet.

In Excel Help, the equations for the computations for CORREL, COVAR and PEARSON are given.

EXCEL DATA ANALYSIS TOOLPAC ROUTINES

There is also a covariance routine and a correlation routine in the Data Analysis tool. In KBA 829208, Microsoft described these Tool Pac routines:

CORRELATION

There is a small difference between the Correlation tool and the Covariance tool that persists in all versions of Excel. The Correlation tool returns a lower triangular correlation table with 1's on the diagonal and correlations off the diagonal. The tool uses CORREL to compute off-diagonal entries and fills those entries with the value (the

numbers, not the function) that is returned by CORREL. (Therefore, if any data entry changes, no entry in the table changes. Contrast this behavior with the behavior of Covariance.)

COVARIANCE

This tool returns a lower triangular covariance table with variances on the diagonal and covariances off the diagonal. Cells on the diagonal contain a formula "=VARP(...)" so that if a data entry changes, the result in the table diagonal also changes, but the other cells do not.

The Covariance tool uses COVAR to compute off-diagonal entries and fills those entries with the value that is returned by COVAR. Therefore if a data entry changes, the off-diagonal entries do not change.

EXCEL 97 AND 2000

REPORTED PROBLEMS

In KBA 215706, Microsoft stated that there was an error in COVAR in Excel 97 that was fixed in Excel 2000. "This behavior occurs because Excel 97 incorrectly uses $1/(1-n)$ instead of $1/n$ in the covariance equation. Excel 2000 correctly uses $1/n$."

McCullough (2000) stated that the CORREL function gave values with substantial error when compared with the NIST biserial correlation values on the StRD univariate data sets. After discussing this with McCullough (2003), my view is that he made a reasonable interpretation of CORREL as described in Help. However the NIST equation is quite different from the CORREL equation as given in Help. This is further discussed in note L.

In KBA 828129 on PEARSON, Microsoft said, "In versions of that are Excel earlier than Excel 2003, PEARSON may exhibit round-off errors. The behavior of PEARSON has been improved in Excel 2003. CORREL has always been implemented with the improved procedure that is now used in Excel 2003. Therefore, if you are using PEARSON for a version of Excel that is earlier than Excel 2003, Microsoft recommends that you use CORREL instead."

RESULTS OF TESTS

There are no specific StRD tables for the testing of covariance and correlation routines. I used the StRD Norris and Longley data sets as a basis for a test. The Norris data set was small, and the reference was the StRD reported coefficient of determination or R^2 value. For the Longley data set, the reference set was the Longley set centered about the means of each variable. The augmented set was the Longley set with 100,000,000 added to each x variable value. This effectively gave 9 significant figures to each variable.

Original Data, N=16

Table 8-2: Longley Data Set Characteristics

Variable	Name	Mean	Original Data L10COV	Augmented Data L10COV
A	Response	65317	1.269	4.455
B	GNP Deflator	101.68	0.974	6.967
C	GNP	387698	0.591	3.004
D	Unemployment	3193	0.534	5.030
E	Military Employment	2607	0.574	5.157
F	Population	117424	1.227	4.158
G	Year	1954	2.655	7.364
	Average		1.118	5.162

Table 8-3: Longley Data Set Reference Correlation Coefficients (r)

	A	B	C	D	E	F	G
A	1.000000						
B	0.970899	1.000000					
C	0.983552	0.991589	1.000000				
D	0.502498	0.620633	0.604261	1.000000			
E	0.457307	0.464744	0.446437	-0.177421	1.000000		
F	0.960391	0.979163	0.991090	0.686552	0.364416	1.000000	
G	0.971329	0.991149	0.995273	0.668257	0.417245	0.993953	1.000000

Table 8-4: COVAR, Excel 2000 LRE Values, Reference Set Versus Original Data

	A	B	C	D	E	F	G
A	16.00						
B	16.00	15.80					
C	16.00	16.00	16.00				
D	16.00	16.00	16.00	16.00			
E	16.00	16.00	16.00	16.00	16.00		
F	16.00	15.68	16.00	16.00	16.00	16.00	
G	16.00	15.53	16.00	16.00	16.00	16.00	16.00

Table 8-5: COVAR, Excel 2000 LRE Values, Reference Set Versus Augmented Data

	A	B	C	D	E	F	G
A	16.00						
B	10.35	10.14					
C	16.00	10.28	16.00				
D	16.00	10.01	16.00	16.00			
E	16.00	10.00	16.00	16.00	16.00		
F	16.00	10.15	16.00	16.00	16.00	16.00	
G	16.00	10.29	16.00	16.00	16.00	16.00	16.00

NewCOVAR (Table 8-6) is the Welford-Kahan algorithm adapted to calculate covariance.

Table 8-6: NewCOVAR, Excel 2000 LRE Values, Reference Set Versus Augmented Data

	A	B	C	D	E	F	G
A	13.01						
B	10.40	10.18					
C	13.23	11/31	13.48				
D	12.47	9.76	16.00	16.00			
E	12.10	9.91	16.00	16.00	16.00		
F	12.99	10.62	16.00	16.00	16.00	16.00	
G	12.91	10.93	16.00	16.00	16.00	16.00	16.00

Table 8-7: CORREL, Excel 2000 LRE Values. Reference Set Versus Original Data

	A	B	C	D	E	F	G
A	16.00						
B	16.00	15.65					
C	16.00	16.00	16.00				
D	16.00	16.00	16.00	15.95			
E	16.00	16.00	16.00	16.00	15.95		
F	16.00	15.64	16.00	16.00	16.00	15.95	
G	16.00	15.22	16.00	16.00	16.00	16.00	16.00

Table 8-8: CORREL, Excel 2000 LRE Values. Reference Set Versus Augmented Data

	A	B	C	D	E	F	G
A	16.00						
B	11.08	15.95					
C	16.00	10.78	16.00				
D	16.00	10.21	16.00	15.95			
E	16.00	9.87	16.00	16.00	15.95		
F	16.00	10.47	16.00	16.00	16.00	16.00	
G	16.00	10.84	16.00	16.00	16.00	16.00	16.00

NewCORREL is the Welford-Kahan algorithm adapted to calculate correlations.

Table 8-9: NewCORREL, Excel 2000 LRE Values. Reference Set Versus Augmented Data

	A	B	C	D	E	F	G
A	16.00						
B	11.16	16.00					
C	13.59	10.42	16.00				
D	12.69	9.88	12.41	16.00			
E	12.14	10.04	12.16	11.74	16.00		
F	12.52	10.24	13.47	12.09	11.71	16.00	
G	13.04	10.35	14.84	12.49	12.15	14.33	16.00

Table 8-10: PEARSON, Excel 2000 LRE Values, Reference Set Versus Original Data

	A	B	C	D	E	F	G
A	16.00						
B	13.77	16.00					
C	16.00	14.84	16.00				
D	16.00	14.13	16.00	16.00			
E	16.00	14.19	16.00	16.00	16.00		
F	16.00	13.55	16.00	16.00	16.00	16.00	
G	16.00	12.31	16.00	16.00	16.00	16.00	16.00

Table 8-11: PEARSON, Excel 2000 LRE Values, Reference Set Versus Augmented Data

	A	B	C	D	E	F	G
A	16.00						
B	1.83	16.00					
C	7.95	1.83	16.00				
D	6.09	1.83	6.22	16.00			
E	5.99	1.83	6.04	4.69	16.00		
F	7.51	1.83	8.47	5.94	6.00	16.00	
G	1.52	2.10	1.53	1.53	1.53	1.52	16.00

Table 8-12: LRE Values, PEARSON, Excel 2000 Reference Set Versus CORREL, Excel 2000 on Augmented Data

	A	B	C	D	E	F	G
A	16.00						
B	11.16	16.00					
C	13.59	10.42	16.00				
D	12.69	9.88	12.41	16.00			
E	12.14	10.04	12.16	11.74	16.00		
F	12.52	10.24	13.47	12.09	11.71	16.00	
G	13.04	10.35	14.84	12.49	12.15	14.33	16.00

CONCLUSIONS

The covariance function (COVAR) is robust against number of significant figures. As shown in the Function Reference, The input range of both the X and Y values is internally centered about the means of the two variables. The divisor of the summed product of the differences of each variable from its mean is n, rather than n-2. Most statistics books take the covariance computation as being divided by n, the number of values. Rather than the number of degrees of freedom.

The correlation coefficient function (CORREL) is not robust against a large number of significant figures. According to the Function Reference, CORREL uses COVAR and STDEV functions, COVAR for the numerator and STDEV on both variables to obtain (as a product) the denominator. STDEV is not robust and is the main contributor to inaccuracies. Also the value is biased to $n/(n-1)$ when X and Y are identical.

The Pearson product moment (PEARSON) is essentially CORREL without the $n/(n-1)$ bias. It however is not robust, and easily produces inaccurate values as the number of significant figures increases. If used, all data should be first centered.

The actual CORREL output is correct, with an LRE value of 15.48.

With reference to the Longley data set, there is a general loss of accuracy between variable B and the other variables. This is due to variable B being much smaller with a smaller variation (L10COV =0.976) than the other variables. There does not seem to be a good solution to increasing the accuracy here, since neither centering nor the Welford-Kahan algorithm (i.e. NewCOVAR and NewCORREL) calculates more accurate values.

The lower LRE values from the Welford-Kahan algorithm (see Note P), is interesting. This shows that what appears to be a better algorithm does not necessarily show up in testing, with greater accuracies. The Welford-Kahan algorithm did give a slightly different reference set (the lowest LRE on the difference was 15.44), but this is not enough to explain the difference in LRE values in the A variable column. Welford's is exact for the $(X-X\text{mean})^2$ term, but the same method adapted to the $(X-X\text{mean})*(Y-Y\text{mean})$ product term is not exact.

Values calculated from the Data Analysis correlation and covariance tools are identical to the values from the CORREL and COVAR functions.

The Excel functions COVAR and CORREL hold up very well to additives. The success for all the variables is due to the algorithm first centering, and then doing the sums of the products. PEARSON does badly because the algorithm as shown in Help is faulty. The correct formula for the Pearson Product Moment Correlation Coefficient is identical with the correlation coefficient as calculated by CORREL.

RECOMMENDATIONS

The PEARSON function becomes inaccurate under the same conditions that the STDEV function loses its accuracy. Numerical differences are due to the differences in the algorithms. Centering of the data is not needed for COVAR and CORREL. For correlations, CORREL is preferred, since CORREL will have a somewhat better

accuracy, while PEARSON will not. For RSQ values, square the CORREL value rather than using the RSQ function.

The L10COV effect on reducing LRE values does occur, but it is hard to make generalizations at which point the value is considered inaccurate, Based on the average of the LRE values across the set, an average below 2 would generally give unacceptable COVAR, CORREL and PEARSON values, but values of 5 and higher give accurate results.

EXCEL 2003 AND 2007

CHANGES FROM EXCEL 2000 IN EXCEL 2003 AND 2007

One of the major changes to many of the functions for 2003 was to go to a two-pass method, which is described above under Univariate Analysis. In KBA 828888, Microsoft said, “Other functions that require a sum of squared deviations about a mean and that have always used the two-pass procedure are CORREL and COVAR. PEARSON and CORREL both compute the Pearson Product-Moment Correlation Coefficient. Both yield the same results in Excel 2003, but PEARSON is implemented with the one pass algorithm in earlier versions of Excel.”

In KBA 828129 on PEARSON, Microsoft said, “The procedure that is used in Excel 2003 uses a two-pass process through the data. First, the sums of X's and Y's and the count of the number of observations in each array are computed. From these, the means (averages) of X and Y observations can be computed. Then, on the second pass, the squared difference between each X and the X mean is found; these squared differences are summed. The squared difference between each Y and the Y mean is found; these squared differences are summed. Additionally, the products $(X - X \text{ mean}) * (Y - Y \text{ mean})$ are found for each pair of data points and are summed. These three sums are combined in the formula for PEARSON.” None of these three sums are affected by adding a constant to each value in the Y array (or the X array), because that same value is added to the Y mean (or the X mean). In the numeric examples, even with a high power of 10 in cell D12, these three sums are not affected and the results of the second pass are independent of the entry in cell D2. Therefore, the results in Excel 2003 are more stable numerically.

For the Tool Pak routines:

Correlation: This tool has not been changed.

Covariance: This tool has not been changed, but the improvement in VARP will fix the errors in values along the diagonal.

RESULTS OF TESTS ON EXCEL 2003

Applying the same testing procedure used above to the table with the largest PEARSON errors shows the improvement in accuracy. Table 7-13 is the results for PEARSON on the augmented data.

Table 8-13: PEARSON, Excel 2003 LRE Values, Reference Set Versus Augmented Data

	A	B	C	D	E	F	G
A	16.00						
B	11.00	15.95					
C	16.00	10.78	16.00				
D	16.00	10.21	16.00	16.00			
E	16.00	9.87	16.00	16.00	16.00		
F	16.00	10.47	16.00	16.00	16.00	16.00	
G	16.00	10.84	16.00	16.00	16.00	16.00	16.00

Since there were no changes to the other functions from Excel 2000, the tests above on CORREL and COVAR are valid for Excel 2003.

MISSING DATA

Empty cells (having the “empty” state) are not detected as an error in the Excel 2003 covariance and correlation functions and routines. When encountered in the computation sequence, the computation element (i.e. count, summation for means, calculation of deviation from mean, product terms, etc.) skips the cell. Consequently the final results will not convey any information that there is missing data. A numerical value will be returned. If the “empty cell” was unintentional, then the result is a wrong number.

The functions and routines will detect non-numeric data, and return an error message.

It is then the responsibility of the user to check the input data ranges for missing values. Excel cannot be faulted regarding wrong answers, if the data range has missing values.

RECOMMENDATIONS

The correlation functions COVAR, CORREL, PEARSON and RSQ are fully accurate to the limits of the IEEE-754 double precision number system. Any perceived inaccuracies are due to the IEEE-754 limits and the basic problems of correlating two number sets that differ greatly in magnitude and have individually small variance.