Journal of Modern Applied Statistical Methods

Volume 5 | Issue 2 Article 29

11-1-2005

Statistical Tests, Tests of Significance, and Tests of a Hypothesis Using Excel

David A. Heiser
United States Air Force, Retired, dah_box1@innercite.com

DOI: 10.22237/jmasm/1162355280

Journal of Modern Applied Statistical Methods
November, 2006, Vol. 5, No. 2, 551-566

1538 – 9472/06/$95.00

Statistical Software Applications and Review
Statistical Tests, Tests of Significance, and Tests of a Hypothesis Using Excel

David A. Heiser
Environmental Management
United States Air Force, Retired

Microsoft's spreadsheet program Excel has many statistical functions and routines. Over the years there have been criticisms about the inaccuracies of these functions and routines (see McCullough 1998, 1999). This article reviews some of these statistical methods used to test for differences between two samples. In practice, the analysis is done by a software program and often with the actual method used unknown. The user has to select the method and variations to be used, without full knowledge of just what calculations are used. Usually there is no convenient trace back to textbook explanations. This article describes the Excel algorithm and gives textbook related explanations to bolster Microsoft's Help explanations.

Key words: Excel, spreadsheets, statistical functions, hypothesis testing, t test

Introduction

Testing any commercial/academic statistically oriented computer program for correctness and accuracy runs directly into the questions, what is correctness and what is accuracy. Unfortunately, the answers are user dependent in the sense that each user has a different answer. The fact is that all commercial/academic software at sometime gives incorrect values, but that doesn't stop users from using it.

"There's a credibility gap: We don't know how much of the computer's answers to believe. Novice computer users solve this problem by implicitly trusting in the computer as an infallible authority; they tend to believe that all digits of a printed answer are significant. Disillusioned computer users have just the opposite approach; they are constantly afraid that their answers are almost meaningless" (Knuth 1998, p229).

David A. Heiser, B.S, University of Wisconsin, M. S., California Institute of Technology, both in Chemical Engineering. He maintains the web page on using Excel in statistics at http://www.daheiser.info/excel/frontpage.html. Email: dah_box1@innercite.com

The question is here, how much of Excel's computed output is believed to be correct and just what is correct?

The EXCEL Spreadsheet Program

Microsoft's Excel spreadsheet program is an inexpensive program for doing many kinds of calculations in business, engineering, and science. Excel has functions and data analysis routines for doing statistical calculations. There are many introductory statistics books that include instructions for solving problems using Excel. Excel also has basic chart and graph capabilities for displaying data and results.

Excel remains very popular, because it allows easy integration with Microsoft's Word and with Microsoft's Access (large business data bases). Results in the form of tables and charts can be easily integrated with Microsoft's PowerPoint presentation software. The pivot table feature as a means of analyzing data is a very popular feature.

Excel's capabilities are limited by the fact that it only does simple statistics. It does not include a lot of additional functions and routines that reflect current commonly used statistical procedures. It was programmed prior to 1992 and version 4.0 in 1994 was the first fully documented version (Excel 1992). It has had essentially no major improvements in statistical capabilities since then. Significant changes

(corrections and improvements) were made for the Excel 1997 and Excel 2003 versions, but the basic module remained the same.

The Computer Environment

It is important for people who deal with numerical computations to understand that the computer works only with a subset of real numbers {IR}. It is a special kind of mathematical object, a field. The computer software uses a different object {IF} to simulate {IR} objects. These objects are called floating point numbers. The object defined by {IF} is a finite subset of {IR}, it is not however, a field (nor any other object that mathematicians commonly define and study) (Gentle, 2004).

In computer software, addition and multiplication of {IF} objects are not associative. The summation in {IF} is not well defined, and usually is taken as a number when its value no longer changes. This no-further-change limit is referred to as being {IF}-convergent, which is different from {IR}-convergent. The harmonic series (sum of $1/i$) in {IR} is divergent, but in {IF}, it is {IF}-convergent. The {IF}-convergent value can be different, depending on how the internal algorithm does associations. The sum of integers is {IF}-convergent, because there is a limit on the size of integers that can be represented as {IF} objects (Gentle, 2004).

The Excel functions and routines handle numbers as the IEEE-754 64 bit standard floating point double precision number. The following are descriptions from KBA 78113:

> "A floating-point number is stored in binary in three parts within a 65-bit range: the sign, the exponent, and the mantissa.

1 Sign Bit	11 Bit Exponent	1 Hidden Bit	52 Bit Mantissa

> The sign stores the sign of the number (positive or negative), the exponent stores the power of 2 to which the number is raised or lowered (the maximum/minimum power of 2 is +1,023 and -1,022), and the mantissa stores the actual number. The finite storage area for the mantissa limits how close two adjacent floating point numbers can be (that is, the precision). (KBA 78113)
>
> The mantissa and the exponent have fixed sizes. As a result, the amount of precision possible may vary depending on the size of the number (the mantissa) being manipulated. Whenever a computation is made (or a value input), the mantissa bits are moved left one at a time and the exponent bits are re-set until the left most bit is a one. Then one more shift is made, transforming this one-bit of information to the hidden bit. Zero bits are added on the right to fill out the 52-bit mantissa." (KBA 78113)

An augmented mantissa of 53 bits corresponds to 15.7 decimal digits. Excel only displays the rounded 15 decimal digits.

> "Every decimal integer can be exactly represented by a binary integer; however, this is not true for fractional numbers. In fact, every number that is irrational in base 10 will also be irrational in any system with a base smaller than 10.
>
> For binary, in particular, only fractional numbers that can be represented in the form p/q, where q is an integer power of 2, can be expressed exactly, with a finite number of bits.
>
> Even common decimal fractions, such as decimal 0.0001, cannot be represented exactly in binary. (0.0001 is a repeating binary fraction with a period of 104 bits)." (KBA 78113)

Errors occur during computer arithmetic {IF} operations.

Round off error.

Results when addition and subtraction are performed. Also occurs in multiplication and division when the sequences involve interchanges between internal 80 bit registers and external 64 bit memory storage. The Excel display also involves another round off.

Overflow and underflow.

Results when the sequence of instructions results in one of the intermediate values either exceeding 1.797693134862315E + 308 (fpmax) or being less than 4.940656458412465E-324 (fpmin). An error return does not always occur. Changing the associations will result in different results.

Quantizing error.

Results when the decimal number cannot be exactly represented by the IEEE-754 binary representation.

The IEEE-754 standard also has an 80-bit floating-point standard. This standard retains the same bit pattern as the 64-bit standard, but extends the mantissa (to the right) an additional 16 bits to a total of 68-bits. Microsoft uses the 80-bit standard for the machine registers that contain the floating-point numbers. At the machine level, computations are done using the 80-bit standard. If however in the sequence of instructions, one of these registers has to be stored in memory, the 80-bit number is rounded to the 64-bit standard and transferred to memory. A multiply-divide sequence that transfers intermediate values to memory will have a different result than one in which the intermediate values are held in the 80 bit floating-point registers. The issue on round-off errors comes from the conversion of the 80 bit number to a 64 bit number.

KBAs 42980, 78113, 145889, 125056 and 214118 are some good sources of information on the {IF} problem. McCullough (1998) also discussed this problem. Knuth (1998) presented the basic theoretical problems of accurately adding, subtraction, multiplying and dividing using floating point numbers as the {IF} object. Higham (1993) also found that there is no universal way to correct for addition (and subtraction) errors in long lists in floating point form.

Algorithms and Computer Programs

This is the area where the mathematics is converted into computer instructions. The general process is to take the mathematics (the equations) and to break the sequences into a series of computing blocks (i.e. subroutines). Then for each of the subroutines, develop (or find in the literature) algorithms made up of fundamental arithmetic type operations (addition, subtraction, multiplication, division, etc) that will perform the desired computations. Subroutines will be written using a computer language such as Fortran, C++, or Visual Basic. The final step is then a conversion (compiling) to a sequence of binary machine instructions (i.e. Intel chip level).

Building a robust algorithm that always gives correct values is not an easy task. For example, take the simple computation of the standard deviation of a list of numbers.

$$\sigma = \sqrt{(\Sigma (x_i - x_{ave}) / (n-1))} \qquad (1)$$

This computation would be done using the calculator formula

$$\sigma = \sqrt{\{ [(n\Sigma x_i^2) - (\Sigma x_i)^2] / [n(n-1)] \}} \qquad (2)$$

with internal summation loops (Knuth, 1998, p 232). This will occasionally require a square root of a negative number, and the overall accuracy is poor. Excel 2000 and earlier versions used this calculator formula to calculate standard deviation values. Excel 2003 uses a two pass method, first calculating an average, then in the second pass calculating deviations from the average, a sum of squares of the deviations and then the standard deviation (KBAs 828888 and 826248). An improved algorithm is Welford's (1962), which is recommended by Knuth (1998). Knuth's form of the algorithm is provided below. Both the mean and the standard deviation are outputted values.

```
DIM Data X(1 to N) As Double
DIM M1, M2 ,S1, S2 as Double
DIM N, K As Integer
M1 = X(1)
S1 = 0
FOR K = 2 to N
M2 = M1 + (X(K)-M1) / CDBL(K)
S2 = S1 + (X(K)-M1) * (X(K) – M2)
M1=M2
S1=S2
NEXT K
AVERAGE = M2
STDEV = SQRT(S2/ CDBL(N – 1) ) )
```

Note: CDBL converts integers to a floating point numbers

Use of the third algorithm substantially improves the accuracy of the result in Excel 2000, but only slightly in Excel 2003. Other statistical computer programs use other algorithms. Maechler (2005) chose West's modification of this algorithm. As he stated, "I'd conclude from Communications ACM, Vol 22, No. 9, page 531, that Welford's algorithm is a bit less accurate than the (very similar) 'West' version, and we (the R developers) should rather implement the latter."

Algorithms sometimes show strange results for an unusual set of input values. For example, enter three identical values, 1E+30, 1E+30 and 1E+30 into Excel cells and do a STDEV function on this range. The result is 1.72368E+14, not zero as expected. Also, do a VAR on this range and 2.97106E+28 will appear.

This raises an important issue. When input of parameter values from one narrow, unusual region of input parameter space results in a wrong output, does one conclude that the computer program should never be used because it returns wrong values?

The Display Of The Result

Within the computer program there are internal subroutines that convert the binary floating point word (64 bits) to a string of ASCII characters (text) which are displayed/printed. The user can (in Excel) chose how the text is formatted as to text type, size, bold, italic, floating point or fixed point and the number of decimals to the right of the decimal point. In Excel there is a default set (Arial, 10, regular), a default cell width of 8.43 points, and the default General format. For numbers from 1 to 0.0001, the General display will show 6 decimal digits. Below 0.0001, a floating point display of 3 digits (plus exponent) will be displayed.

There have been articles published criticizing the accuracy of computer software based solely on the default display (e.g., Altman 2002, Hilbe, 2002, McCullough, 1998, 1999, McCullough & Wilson, 1999, 2000, 2004; Knŭsel, 1998, 2003).

Methodology

McCullough (1998, 1999) pioneered some of the basic methods of conducting tests on software. He used the NIST suite of data-bases with known statistics to test several software programs. His two articles are good background and methodology sources.

Testing methods

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. For other programs, there may be all kinds of decision trees and selections to arrive at the test objective or method to be tested.

What is the function/routine actually doing? In most cases, the developer says very little regarding the specifics of what the program does, but a great deal is said on marketing (selling) how good and comprehensive is the program. For proprietary reasons, of course, very little should be said. For that reason, some testing has to be done to find out just exactly what is being calculated, how to get as many digits as possible, and to find some boundaries on the ranges of input parameters. This is exploratory testing.

The next level is accuracy testing. 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.

Standard values of summary statistics from a data set may come from several sources.

1. Theoretical values manually calculated or selected (by theory) 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: A).
2. Values calculated by an external software program, chosen to be the reference (Method: B).
3. Data and values published as part of a standard. (Method C).
4. Comparing the results from many different software programs on the same data set and deciding on "correctness" (Method D). Altman and McDonald (2000).

The NIST Tests

The National Institute of Technology (NIST) established datasets for software tests, the StRD series (NIST nd).

> "For all datasets multiple precision calculations (accurate to 500 digits) were made using the post-processor and FORTRAN subroutine package of Bailey (1995, available from NETLIB). Data were read in exactly as multiple precision numbers and all calculations were made with this very high precision. The results were output in multiple precision, and only then rounded (without error) to fifteen significant digits. These multiple precision results are an idealization. They represent what would be achieved if calculations were made without round-off or other errors. Any typical numerical algorithm (i.e. not implemented in multiple precision) will introduce round-off error, and will produce results that differ slightly from these certified values." (NIST, nd)

The NIST data sets covered univariate analysis, linear regression, non-linear equation fitting, ANOVA and correlations. This has been the essential test method (method C) to test Excel. McCullough (1998, 1999) pioneered the basic method of conducting tests on software using the NIST test sets. McCullough and Wilson (1999, 2000, 2004) also presented a series of papers on tests made on Excel using the NIST and other test data .

Other Previous Excel Tests

Some of the early testing (Excel 1995) was done by the Center for Information Systems Engineering, (Britain) in 1999 (CISE 27/99). They used the IMSL Fortran 90 Math/Library (version 3.0) provided by the Digital Equipment Corporation to do testing (Method B).

A number of email messages, web site reports (papers), and discussions on the newsgroups and on the statistical lists (since 1998) described tests on some of the Excel functions and routines. These included cases where a particular (real) data set, when analyzed using Excel, gave results different from some other software package. Most of these were casual tests, based on a particular data set.

Significance Test Methods

The NIST data sets and their computed statistics were not useable on the family of significance tests in Excel. NIST did not provide paired or dual data sets for testing significance test functions/routines. The literature does not report on specific testing of Excel significance test functions and routines. Therefore, test data sets for testing the Excel family of significance tests had to be built, and ways to arrive at accurate statistical values found

Because the outputs from some of the significance tests are p values, a set of Visual Basic statistical distribution functions provided by Smith (2002) were used to calculate accurate reference p values. The Excel distribution functions are not accurate enough to be used to obtain accurate p values.

Two approaches were taken, one of exploratory testing to identify just what the function was returning (e.g., the proper tail area). The other was to do accuracy testing. This required the development of more extensive data sets to stress the functions/routines.

The NIST approach was to use several types of test data sets. One of these types was to build patterned data tables of data. A patterned number can be considered as having a whole number part and a fractional part where the

numbers to the right of the decimal point is the fractional part. A patterned data table has patterned numbers all with the same whole number, but with different fractional values. For the NIST SmLs01 to SmLs09 data sets, the fractional part had specific alternating values (0.3 and 0.5 or 0.2 and 0.4), and then with one odd value for each set, gave a data set with theoretical, precise means, variances and standard deviation values. By increasing the magnitude of the whole number from 1 to 1E+09, and by changing the size of the set, the overflow effect on floating point number computations and algorithms could be determined.

The NIST approach to the SmLs sets suggested ways to build test data sets with accurate statistics to test the Excel family of significance tests. The theory behind it comes from the basic way numbers are represented in Excel.

In terms of floating point numbers, a larger whole number part of the patterned number pushes the mantissa bits (these are on a number base of 2, not on a number base of 10) off the right end, characteristic of overflow. This overflow of floating point numbers is one of the causes of errors. However, there are other causes of errors that are not brought out by the use of patterned numbers, and other methods have to be used. Good algorithms are those that minimize the overflow effect. The charts in Heiser (2005) show the loss of accuracy of many Excel functions due to this type of overflow.

Measures Of Accuracy - Log Relative Error (LRE)

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 in his 1998 paper. The LRE value represents a measure of how many significant (accurate) digits (decimal) there are in the output parameter values.

$$LRE = -LOG_{10} (abs (CV-RV) / RV))$$

CV is the computed value and RV is the reference or true value. LRE values vary from 0 to 15 on the McCullough scale. 15 can be considered as an exact match.

LRE values from the statistical distributions present problems, because of the 9's problem. Here, a leading sequence of 9's really are leading zeros, and should not be considered as significant digits, but mathematically they are. Excel computes p values above 0.5 as 1 minus the corresponding below 0.5 value, for all symmetric probability distributions. Consequently, p values above 0.5 have uncertain accuracies, depending on the user's view. Smith's (2002) distribution functions calculate p and q values by separate algorithms.

The LRE values approximate the number of accurate digits in the Excel cell value, independent of how it is displayed. For the floating point form, (select Format→ Cells→ Number→ Scientific→ Decimal Palaces→ 14) it approximately represents the number of accurate digits, including the digits to the left of the decimal point, and the digits to the right of the decimal point.

Results of Tests

This study examined the errors from the Excel VAR algorithm and Welford's algorithm on a patterned data set. In this case, two sets of random fractional numbers, one uniform u(0-1) and the other normal n(0,1) with 1001 values of each set were generated in a column (Please note that for all test data sets with random numbers, Marsaglia's MWC256 RNG, Marsaglia (1995, 2002) was used. For random normal, Smiths's (2002) precise inverse normal function was used). The variance value of the base case from either of the two functions was the identical. Whole number sets (from 1 to 1E+15) were added forming 15 additional columns. Variances from each function were then calculated. Figure 1 shows the result.

Given the nature of the input data and the basic structure of a patterned number in terms of the decimal system, the data from a good algorithm should closely follow a straight line from 16 on the y axis to 16 on the x axis.

The Excel 2003 algorithm, although an exact algorithm, shows some unexpected behavior in the region below an exponent of 8. This behavior generally occurs also for other Excel functions when the whole number is less than 1E+08. The inaccuracies at the right end

are expected. Welfords's algorithm in general is close to the expected line and shows consistent behavior, typical of a good algorithm.

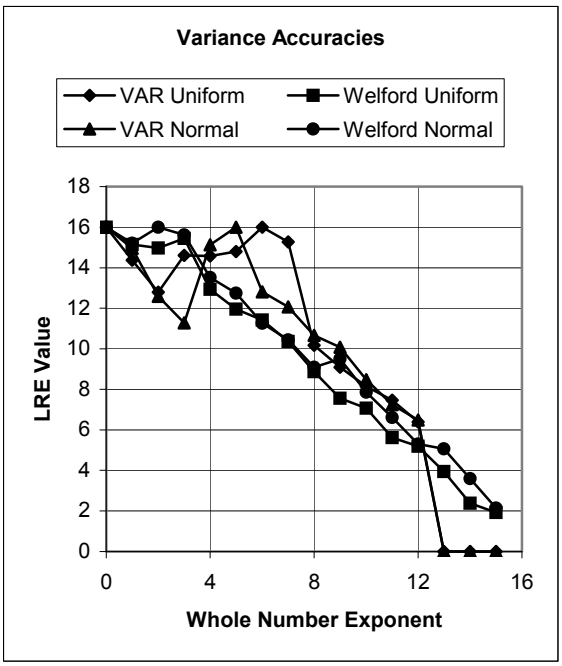

Figure 1: Comparison of Algorithms

The Excel Significance Test Functions And Routines

Excel 2003 provides 80 direct functions and 19 Data Analysis routines that can be used in statistical data analysis. Only a part of the available functions and routines are directly applicable to tests of significance and hypothesis testing. The functions and routines useful for significance testing are:

CHITEST - This is a Chi-square Goodness-of-Fit test for grouped data. It does not support general Chi Square tests on variances. The test will only work on 2 way contingency tables. The test cannot be applied to single lists of observed and expected values. The first input, actual range is the range of the observed values, as a 2-way contingency table. The second input is expected range, the range of a separate contingency table giving the expected values.

FTEST - Returns the one-tailed probability value of an F test on two separate ranges of data. The ranges may be of different lengths.

TTEST - Returns the probability value of a t test on two separate data sets. Function allows for 1 or 2 tail tests, paired data and equal-unequal variances. The function has two parts internally, one to calculate a t value from the two separate data sets, and the other to calculate internally a p value from the t value.

ZTEST - Returns the two-tailed probability of a normal distribution z test on a range of data with respect to a known population mean and standard deviation. If the standard deviation field is left blank, the routine used the standard deviation of the data. The function has three parts internally:

1 To calculate a mean value (and a standard deviation) from the input data set.

2 To calculate z = [(input mean value) – (data set mean)] / [(data set standard deviation or input standard deviation) / Square Root (size of the data set, n)].

3 To calculate a p value from NORMSDIST(z).

All of the other Excel functions can be used to build up intermediate values for significance test inputs. They can also be used along with new VBA functions and subroutines to build new significance tests beyond the limited capability of Excel.

Data Analysis Routines

These are routines called by selecting the *Tools* menu and then selecting *Data Analysis* and then selecting one of the listed routines.

F-Test Two-Sample for Variances
t-Test: Paired Two Sample for Means
t-Test: Two-Sample Assuming Equal Variances
t-Test: Two-Sample Assuming Unequal Variances
z-Test: Two-Sample for Means:

After inputting the requested data, they return a table.

Tests on the Accuracies of Functions and Data Analysis Routines

The CHIDIST, FTEST and TTEST functions were tested. There were differences found between the results of these tests for Excel 2000 and Excel 2003. The Excel 2000 tests show relatively low LRE values. As explained by Microsoft in KBA 828888, the problem was the low accuracy of the VAR and STDEV functions that were used inside the routines. Rather than take up a great deal of space to show both 2000 and 2003 outputs, only the Excel 2003 values are shown in the following tables.

There were 4 data sets used for testing as follows:

Set 1 (columns A and B) represented paired data, integers with blank spaces.

Set 2 (columns C and D) represented unequal length data from two different populations. Integers.

Set 3 (columns E and F) represented patterned data of two samples from one population with equal sample sizes. The whole number was 1000, and the fractional numbers were uniformly distributed (0-1) random numbers.

Set 4 (columns G and H) represented a variable length set (up to 2000). The first column represented the control data set, and the second column represented the treatment data set. The base case was where the numbers in both columns were all random normal (0,1) z values from one population. Whole numbers were added as described previously.

Testing The Difference Between Variances

CHITEST

Tests indicated that the Excel algorithm in the CHITEST function is the correct one. Errors occur from errors in the inputed expected values table and in the CHIDIST function. CHITEST returns correct values if the Expected Values table is correct.

FTEST

The function description (Excel, 1992) suggests that the FTEST function just computes the ratio of two variances where the variances come from the VAR function. Neither Excel Help nor the KBAs provide any additional information. The VAR function holds up well against overload as shown in figure 1, but does introduce some error.

Given the ratio, the FINV function then was used to arrive at a p value. The F distribution FINV generally has p value accuracies above an LRE value of 8, over the entire range of input parameters (see Heiser, 2005) for specific details. The output then of FTEST should be an accurate p value with at least 8 accurate decimal digits. The actual output for data set 2 indicates that FTEST returns wrong values.

Table 1: FTEST Function Response

Cell Entry	=FTEST(C,D)
Returned Value	0.9425381810184540
Correct Value	0.481410961628470

FTEST outputs an incorrect p value, corresponding to a two-tail test. The problem is with Microsoft.

In Excel (1992), the function description says, “Returns the results of an F-test. An F-test returns the one-tailed probability that the variances in array 1 and array 2 are not significantly different”.

In Excel Help (2006), “Returns the result of an F-test. An F-test returns the two-tailed probability that the variances in array1 and array2 are not significantly different. Use this function to determine whether two samples have different variances.”

The standard for the F test on a ratio of variances is the one tailed test. It is a test on all values of the ratio from 0 to the critical value. On this basis, the only valid test is the one–tailed test. The workaround here is to always divide the FTEST p value by 2 to get the correct q value of the right tail. This has been reported before.

Test On The Data Analysis F-Test: Two-Sample For Variances:

Here Excel returns an accurate value.

Table 2: Excel Data Analysis Routine Output, Actual Excel Output

F-Test Two-Sample for Variances		
	C	D
Mean	1000.503767	1000.696727
Variance	0.092055155	0.090461689
Observations	30	30
Df	29	29
F	1.017614821	
P(F<=f) one-tail	0.481410962	
F Critical one-tail	1.860811434	

The true p value is 0.481410961628470. The Data Analysis F test on two variances gives the correct p value (excluding the argument on the correctness of all displayed digits). Differences are only due to the inaccuracies in FINV (the function that uses the df and F ratio values to arrive at a p value)

Testing the Difference Between Mean Values the Basic Problems and Solutions

There are three possible situations or problems here with tests on the differences in means.

(1) Dependent, Paired values,

(2) Independent, Two sample sets, each coming from different (or the same) population with possible differences in means, but both populations having the same variance

(3) Independent, Two sample sets, each coming from different populations with different variances (The Fisher-Behrens problem).

These are the three classical situations, which require different test methods.

Excel provides a function (TTEST) with 3 options and three Data Analysis routines for statistical solutions for the basic problem. The questions here are just what do these functions and routines do, and do they compute the statistics correctly in terms of theory, and are the results numerically accurate. Other concerns include: how robust are they on non-normal data, how stable are the results in terms of type I error rates, and what the power is.

In traditional statistics, the three possible situations are considered as separate, important classical problems for analysis in introductory statistics. In introductory statistics, the assumption of normality is made, and this results in a simplification of the statistical tests. The test is usually put in terms of a test of a hypothesis. The discussion below is based on the traditional tests using the t distribution and the assumption of normality.

The paired values (or dependent data values) solution problem (1) is straightforward, and is given in textbooks. The test is to determine if the sum of the differences between each pair is zero or is some preset difference, depending on the hypothesis made.

For problems (2) and (3), the test is on the differences of the means, using a joint measure of variation from both samples. Problem (2) where the variance does not change and approximately equal sample sizes are involved has a very robust t test solution under sample departures from normality. However, if the variances are not truly equal, and substantially different sample sizes are involved, the normal t test solution looses its robustness and the true alpha may be quite different from the selected alpha. The third problem is the Behrens-Fisher problem, which does not have a direct theoretical solution.

The Behrens-Fisher Problem

For the Behrens-Fisher problem, there is no uniformly most powerful (unbiased) test for

all sample sizes. There are several approximations found in textbooks and in the literature, and this complicates the assessing of Excel's accuracy on problem (3). This impacts the decision to fault Excel or not. Sawilowsky (2002) is an excellent review of the attempts to come up with more exact solutions since 1929.

Fisher's Solution Of The Behren's Problem

> "For samples from a single population, the effect of eliminating the unknown variance σ^2, by Student's method, on the distribution of the error of the mean, is to replace, in the specification of this error,
>
> $$\sigma * x / \sqrt{N}$$
>
> where x is normally distributed with unit variance, but σ is unknown by
>
> $$s * t / \sqrt{N}$$
>
> where t is distributed in Student's distribution, for the appropriate number of degrees of freedom N, and s is the estimate of σ available from N degrees of freedom:"
>
> For two samples from populations having a common mean, the deviations will be independent, and the data will supply values s_1, based on n_1 degrees of freedom, and s_2 based on n_2. The difference between the observed means is the sum (or difference) of the two deviations from the true mean, so that on the null hypothesis considered, namely that the two populations means are equal, we have
>
> $$x_1 - x_2 = (s_1 * t_1 / \sqrt{n_1}) - (s_2 * t_2 / \sqrt{n_2})$$
>
> where t_1 and t_2 are distributed independently in the two distributions.
>
> If the frequency is small, such as 1%, that the expression on the right, which has a known distribution, for the observed values s_1 and s_2, shall exceed the observed difference in the sample means, this difference may be judged significant." (Fisher 1973, p. 98).

This is the same as the confidence interval method described in Schenker and Gentleman (2001), where the s values are population values and the t values are z values. Fisher's method does not lend itself to a direct solution and is not referred to as a solution in the literature.

The Welch-Aspin-Satterthwaite Solution

The Welch-Aspin-Satterthwaite solution is a solution to the Behrens-Fisher problem. It evolved over the years from Satterthwaite's ideas in 1941 to Welch's ideas in 1937 -1949, with Aspin's inputs during 1948-1949. It is commonly referred to as the Aspin-Welch test or the Welch test in research papers. However, some statistics textbooks authors (i.e. Moore & McCabe, 2003) ignore this and use "pooled df for this test, also known as the computer solution". There is no consistency in the literature between the names or terms used and which of six computational methods it applies to.

One of the inherent problems with the Welch-Aspin-Satterthwaite approximate solution is that it is not robust to departures from normality. Sawilowsky (2002), stated, "I would be remiss if I failed to note that numerous Monte Carlo studies have shown that the nonparametric Wilcoxon Rank Sum test can be three to four times more powerful in detecting differences in location parameters when the normality assumption is violated….Therefore the Wilcoxon procedure should be the test of choice".

However this does not resolve the fundamental problem as to whether the difference should be determined based on the medians (Wilcoxon) or on the means (Welch). The predominate applications in psychology and related behavioral research are based on the difference in means, the standard error of the means and on effect size. There is little concern about non-normality and equality of variances. Effect sizes are more important than p values. (Sprinthall, 2000, Kline, 2004)

Excel does not provide the Wilcoxon Rank Sum Test, which can be considered a fault in Excel. For samples that have large differences in sample sizes and have assymmetry, the Balkin and Mallows (2001) approach should be considered.

The Six Solutions

The range of possible solutions to the three situations identified above has to be limited specifically to what Excel has provided. Within the context of what was discussed above, there are 6 possible solutions to the Behrens-Fisher problem.

In general, the p value (compared to the alpha value) comes from the t distribution, and therefore for each problem, a df value and a t value has to be calculated. A decision also has to make, on whether a single tail or a two-tail test is required.

Methods to obtain a t value from the difference between the two means are listed in table 3. There are others such as the Score statistic that are not considered here, because they are not found in or introduced in introductory statistics textbooks.

Table 3: Combined Variance Measures

t-Value Method	Term 1	Term 2	Common names
1	var_1 / n_1	var_2 / n_2	Un-equal variances
2	var_{pooled} / n_1	var_{pooled} / n_2	Equal variances, pooled t test
3	$var_1 / (2n_1 + 1)$	$var_2 / (2n_2 + 1)$	Fisher's 1939 form
4	$var_1 * (n_1 -1) / (n_1^2 - 3n_1)$	$var_2 * (n_2 -1) / (n_2^2 -3n_2)$	Fenstad's Statistic
5	$var_1 * (n_1 -1) / n_1^2$	$var_2 * (n_2 -1) / n_2^2$	Wald Statistic

The pooled variance in method 2 is:

var_1 = Variance of Sample 1

var_2 = Variance of Sample 2

$var_{pooled} = ((n_1 -1) * var_1 + (n_2 - 1) * var_2) / (n_1 + n_2 -2)$

The combining of terms 1 and 2 for a t value are as follows:

Methods 1-4:

t value = Difference in Means / Square Root (Term 1 + Term 2)

Method 5:

t value = $(\text{Differences in Means})^2$ / (Term 1 + Term 2)

Currently, only t-value methods 1 and 2 are considered. The different degrees of freedom used are given in table 4.

Table 4: Degrees-of-Freedom Values Used In The Tests

df-Method	Used on Problems	df value used to obtain the t distribution p value
1	(1)	= n-1
2	(2)	$= n_1 + n_2 - 2$
3	(3)	= Smaller of either $n_1 - 1$ or $n_2 - 1$
4	(3)	= Welch df

This then gives six ways to calculate a p value, as shown in table 5.

Table 5: The Six Calculation Combinations for Problems (2) and (3)

Calculation	df Method	t Value method
1	2	1
2	3	1
3	4	1
4	2	2
5	3	2
6	4	2

Calculation 3 is generally referred to as the Welch test.

The maximum power here for any of the calculation methods is at sample sizes related to the ratio of the known variances of the samples.

κ = variance population 2 / variance population 1

Where the optimum sample sizes (n1 and n2) come from the following equation:

$$n1 / (n1+n2) = 1 / (1+ \sqrt{\kappa})$$

However, the local optimal design is sensitive to the misspecification of the κ value. (Dette & O'Brien, 2004)

The Welch df Value (df method 4)

$$u_1 = (s_1 * s_1) / n_1$$
$$u_2 = (s_2 * s_2) / n_2$$

$$df = (u_1 + u_2)^2 \ / \ [(u_1^2 / (n_1 - 1)) + (u_2^2 / (n_2 - 1))]$$

There are other textbooks and statistics course handouts that give a different formula and also may call it by other names.

Figure 2 shows how the Welch df value varies as the ratio of the variances varies. It is the factor that when multiplied by the df value of one sample (i.e. n1-1) gives the Welch df value.

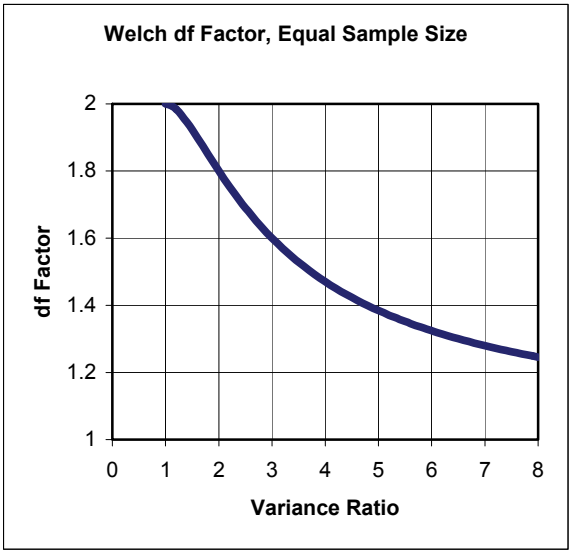

Figure 2: Welch df Factor Changes

Values between 0 and 1 are a mirror image of the values from 1 to infinity, with the x axis values the reciprocals of the x axis values greater than 1. When the variance ratio is 1, the pooled df value is equal to the df method 2 value. As the ratio increases, the pooled df value becomes asymptotic to the df method 3 value.

For example, given equal samples of 30, the F test would probably indicate that variance ratios greater than 2, would indicate a high probability of the variances being unequal. One could conclude then a factor of 1 would be appropriate. However, the Welch-Aspin-Satterthwaite df gives a more conservative estimate that in a sense, compensates for the fact that it is not truly known that the variances are equal.

There have been many articles over the years that point out that if the F test is used to decide on equal/unequal variances at an alpha level, and then do the t test at the same alpha level, there is a subsequent loss of control of the Type I and Type II error rates (e.g., Sawilowsky, 2002).

There are three views regarding the actual Welch df value to be input to the t distribution. The calculated Welch df value is not an integer. The options are to truncate the computed df value to an integer, round to an integer, or interpolate (in tables) to obtain a value for a fractional df. Moore and McCabe (2003) recommended that interpolation be used when only tables are available. Most software routines that calculate the t distribution p value require that the df value be an integer, although the basic computing algorithm will take fractional df values. Excel's t distribution functions will only allow integer df values to be entered.

The Common Textbook Df Value

For unequal variance problems, df method 3 corresponding to calculation 2 is usually given. This results in a conflict here, because Excel follows df method 4.

Best and Rayner (1987) identified four t statistic measures that should be considered:

(V) The common statistic:

(W) The Wald statistic:

(L) The likelihood statistic:

(S) The score statistic:

The common statistic is (V) which corresponds to calculation 3. Best and Rayner (1987) defined the other three (W, L and S), but concluded that for their n1=4 and n2=8 sample sets (from Monte Carlo sets), the power of the test for differences was about the same.

Best and Rayner (1987) defined calculation method 3 as the V statistic They found that calculation 3 gives results that closely follow the preset alpha value, whereas calculation 4 results vary considerably from the designated alpha value when the population variance ratio departs from 1. The V statistic was their choice, because if can be used for both tests involving equal and unequal variances.

Some Textbook Directions

Moore and McCabe (2003) suggested the use of calculation 4 (e.g., Table 6) for equal variances and calculation 2 or 3 for unequal variances. Calculation 3 is preferred for unequal variances. Larson and Farber (2003) said to use calculation 4 for equal variances and calculation 2 for unequal variances. Triola (2001) said to use calculation 4 for equal variances and calculation 2 for unequal variances. Lind, Marchal, and Mason (2001) said to use calculation 4 for equal variance (they do not say anything about unequal variances). Pelosi and Sandiffer (2000) said to use calculation 4 for equal variances and calculation 3 for unequal variances. Levine, Berenson, and Stephen (1999) said to use calculation 4 for equal variances. Unequal variances were not covered. In Sprinthall (2000) the equal/not equal variance issue is never brought up. The standard error of the difference in means is from calculation 4.

The general consensus among textbook authors is to use calculation 4 for equal variances, because it is based on accepted practice. Calculation 2 is more frequently recommended than calculation 3 for unequal variances. In some textbooks, the distinction between equal and unequal variance is not made and calculation 1 is given for all tests on two means from independent samples. This suggests there is a wide range of practices, all derived from whatever was said in the textbook used in the course.

In doing the calculations for tables covering the six methods, calculation method 2 gives higher p values than calculation method 3. Consequently using textbook recommendations may not be the best solution method. They also may lead to false claims about the accuracy of Excel's TTEST and Data Analysis routines.

The Best Approach

In applied studies and research, the current view is that the real problem is that both a shift in location and a change in scale occur simultaneously when a treatment is applied. Consequently both a change in the mean and a change in variance occur. The occurrence of a change in variance without a change in means or a change in means without a change in variance is very rare (Sawilowsky 2002). The third problem then is the main view when dealing with real data.

If the assumption of normality is valid, then the best method is the V test or calculation 3 for all tests on the difference in means, regardless if the variances are equal or unequal. If the test is not a zero difference, but a test on a predetermined (theory) difference (d), then the non-central distribution has to be used rather than the central t distribution. (Steiger & Fouladi, 1997) Excel only has the central t distribution, and therefore Excel cannot be used to test for d.

Computed Reference Values

Computed reference values from each of the six methods for each of the four reference data sets were calculated as a means of finding out which of the methods are used in Excel.

Significance Test Functions and Routines.

The Excel TDIST function (which uses the BETADIST function to derive p values) appears to be used in all cases. There are errors in BETADIST that carry over to the problem solution. An analysis of these function errors and inaccuracies are in Heiser (2005).

Fisher's Solution

Fisher's equation obviously represents confidence intervals, but the signs are a problem. If the left hand sign is taken as a plus (adding two confidence intervals), it is possible to obtain a p value, when the sum of the confidence intervals equals the difference in means. The

theoretical problem is, should the non-central t distributions be used (see Steiger & Fouladi, 1997).

Tests on the TTEST Function.

The TTEST function has three options corresponding to the three possible solutions to differences in means, as discussed above.

Option 1: Paired Sample

The literature has commented on this test, primarily on its failure to give correct values when a BLANK occurs in a cell (indicating a missing value). Both the Excel 2000 and 2003 versions have this problem. Therefore, it is important to never have blank cells in the input ranges.

KBA 829252 describes the odd behavior of TTEST when there is missing data.

When there is no missing data, TTEST returns correct values. The algorithm is correct. The main contributor to errors is the inaccuracy in the BETADIST function that is used to obtain t distribution p values.

Option 2, Two-Sample Equal Variance

Returns correct values. The algorithm is correct. The main contributor to errors is the inaccuracy in the BETADIST function that is used to obtain t distribution p values.

Option 3, Two-Sample Unequal Variance.

Uses calculation 3, the Welch-Aspin-Satterthwaite solution to the Fisher-Behrens problem. Returns correct values. The main contributor to errors is the inaccuracy in the BETADIST function that is used to obtain t distribution p values.

Comments on TTEST

The primary source of errors in TTEST is that from BETADIST. The algorithms used for this function are poor, and consequently often give inaccurate results (see Heiser, 2005). It is not unusual to get LRE values down to 4 with actual data from TTEST because of this problem. Tests on different test data sets generally return p values with LRE values in the 7-10 range. Microsoft has no plans to fix BETADIST, so the use of TTEST will always have this uncertainty.

A fix for this problem is to download an accurate beta distribution function as a vba module or as an *.xla addin from another source. Most t distributions in add-ins or modules are blocked against non-integer df values. Non-integer df values are required for the Welch solution. The relationship (from Abramowitz and Stegun, 1963, eq. 26.5.27) is

$$1 - \text{tdist}(t, df) = \text{Beta}(X, A, B)$$

where

$$A = df / 2$$
$$B = 1 / 2$$
$$X = df / (df + t * t)$$

Beta is the cumulative or incomplete beta function

You will have to modify the left hand side to get the correct one or two tailed probabilities. Smith (2002) has accurate beta distribution vba functions. However, his t distribution functions are restricted to integer df values.

Tests on The Data Analysis Routines

These are routines from the Tools → Data Analysis menu:

t-Test: Paired Two Sample for Means
t-Test: Two-Sample Assuming Equal Variances
t-Test: Two-Sample Assuming Unequal Variances
z-Test: Two-Sample for Means:

They are programmed macros written prior to Excel 4, are not in vba and never have been fixed. Microsoft has issued KBA's on the problems with these macro's but has never fixed the problems.

A consistent error with all four of these routines is the output table where the cells:

P(T<=t) one-tail
t Critical one-tail
P(T<=t) two-tail
t Critical two-tail

appear. The T and t relationships in the first and third cell are usually wrong in terms of the

values given in the cells to the right of this group. KBA 829252 talks about one instance of this problem.

The P(t<=T) two-tail statement is in error. It should be P(T=t) two-tail. A two tailed test in regard to a hypothesis is a test on a null hypothesis of equality. The alternate hypothesis is T≠t. Microsoft is wrong in their help narrative.

t-Test: Paired Two Sample for Means

The Data Analysis Macro uses the TTEST function with option 1, and as a result, the p values from the two are the same.

However there is a difference when blank cells occur. KBA 829252 describes what happens. "First, this Analysis ToolPak tool counts the number of subjects with Before measurements and the number of subjects with After measurements. If these totals are different, you receive an error message and this Analysis ToolPak tool does not continue." Therefore, this routine should not be used when there are missing values in the data.

t-Test: Two-Sample Assuming Equal Variances

The Data Analysis Macro uses the TTEST function with option 2, and as a result, the p values from the two are the same.

t-Test: Two-Sample Assuming Unequal Variances

The Data Analysis Macro uses the TTEST function with option 3. However, there is a difference here. The macro takes the computed Welch df value and converts it by rounding to an integer. This integer value then goes into BETADIST and comes out with a p value different from that coming out of TTEST, option 3. The p values are different here, so in a sense, the Data Analysis macro is in error. For correct Welch p values, fractional df values must be retained. Therefore, this macro gives inaccurate results.

z-Test: Two-Sample for Means

This routine uses the normal distribution function NORMDIST which has LRU accuracies of 7 or more in the z range of –3 to –5, and 12 or more outside of this range. Some tests indicated no algorithm problems, and output p value accuracies corresponding to the NORMSDIST accuracies. However again the P(t<=T) two tail statement in the table is in error. It should be P(T≠t) two-tail.

Excel 2007 (Formerly Excel 12)

This is the new version that will be available in 2007 to work with Windows Vista. Microsoft has made no changes to Excel 2003 in the statistics area for the 2007 version (Gainer, 2006)

References

Abramowitz, M. & Stegun, I. A. (1964). Handbook of mathematical functions with formulas, graphs, and mathematical tables. *National Bureau of Standards, Applied Mathematics*, Series 55.

Altman, M. (2002). A review of JMP 4.03 with special attention to its numerical accuracy. *The American Statistician, 56*(1).

Altman, M. & McDonald, M. P. (2000). *The robustness of statistical abstractions: A look under the hood of statistical models and software.* http://data.fas.harvard.edu/numerical_stability/g3009.pdf

Balkin, S. D. & Mallows, C. L. (2001). An adjusted, asymmetric two-sample t test. *The American Statistician, 55*(3), 203.

Best, D. J. & Rayner, J. C. W. (1987). Welch's approximate solution for the Behrens-Fisher problem. *Technometrics, 29*(2), 205-210.

Cook, H. R, Cox, M. G., Dainton, M. P., & Harris, P. M. (1999). *Testing of spreadsheets and other packages used In metrology, testing the intrinsic functions of Excel*. Center for Information Systems Engineering, NPL Report CISE 27/99.

Dette, H. & O'Brien, T. E. (2004). Efficient experimental design for the Behrens-Fisher problem with application to bioassay. *The American Statistician, 58*(2), 138.

Microsoft Excel function reference for Excel version 4.0 (1992). Manual published by Microsoft Corporation.

Gainer, D. (2006). Email message from Gainer, D, Group Program Manager for Microsoft Excel. See http://blogs.msdn.com/excel/archive/2005/09/23/473185.aspx

Heiser, D. A. (2005). Microsoft Excel 2000 and 2003 faults, problems, workarounds and fixes. http://www.daheiser.info/excel/frontpage.html .

Fisher, R. A. (1973b). *Statistical methods and scientific inference* (3rd Ed). New York, N.Y.: Hafner Press.

Gentle, J. E. (2004). Courses in statistical computing and computational statistics. *The American Statistician, 58*(1), 2-5.

Higham, N. J. (1993). *The accuracy of floating point summation*. Department of Mathematics, University of Manchester Press.

Hilbe, J. B. (2002). Section editor's notes. *The American Statistician, 56*(2), 148.

Microsoft knowledge base articles. http://support.microsoft.com/kb/…..the number …./en-us

Kline, R. B. (2004). Beyond significance testing, reforming data analysis methods in behavioral research. Washington, D.C.: American Psychological Association.

Knüsel, L. (1998). On the accuracy of statistical distributions in Microsoft Excel 97. Computational Statistics and Data Analysis, 26, 375-377.

Knüsel, L. (2003). *Numerical accuracy of distributions in statistical packages*. Paper presented at the International Statistical Institute in Helsinki (1999), 147-148.

Knuth, D. E. (1998). *The art of computer programming (vol 2.). Semi-numerical algorithms* (3rd Ed.). Addison-Wesley

Larson, R. & Farber, B. (2003). *Elementary statistics: Picturing the world* (2nd Ed.). Upper Saddle River, NJ: Prentice-Hall.

Levine, D. M., Berenson, M. L., & Stephan, D. (1999). *Statistics for managers using Microsoft Excel* (2nd Ed.). Upper Saddle River, NJ: Prentice-Hall.

Lind, D. A., Marchal, W. C., & Mason, R. D. (2002). *Statistical techniques in business and economics* (11th Ed.). New York, N.Y.: McGraw-Hill-Irwin.

Maechler, M. (2005). Welford's algorithm. Email: Seminar fuer Statisk, ETH-Zentrum. Zurich, Switzerland.

Marsaglia, G. (1995). *The Marsaglia random number CD-ROM with the diehard battery of tests of randomness.* Available at www.stat.fsu.edu/diehard.

Marsaglia, G. (2003). Random number generators. *Journal of Modern Applied Statistical Methods, 2*(1), 2-13.

McCullough, B. D. (1998). Assessing the reliability of statistical software: Part 1. *The American Statistician, 52*(4), 358-366.

McCullough, B. D. (1999). Assessing the reliability of statistical software: Part 2. *The American Statistician, 53*(2), 149.

McCullough, B. D. & Wilson, B. (1999). On the accuracy of statistical procedures in Microsoft Excel 97. *Computational Statistics and Data Analysis, 31*, 27-37.

McCullough, B. D. & Wilson, B. (2000). On the accuracy of statistical procedures in Microsoft Excel 2000 and Excel XP. Computational Statistics and Data Analysis, 40(4), 713-721.

McCullough, B. D. & Wilson, B. (2004). On the accuracy of statistical procedures in Microsoft Excel 2003. *Computational Statistics and Data Analysis.*

Moore, D. S. & McCabe, G. P. (2003). *Introduction to the practice of statistics* (4th Ed.). New York, N.Y.: W.H. Freeman and Company.

National Institute of Science and Technology. Standard Data. http://www.itl.nist.gov/div898/strd/general/dataarchive.html

Pelosi, M. K., & Sandifer, T. M. (2000). *Doing statistics for business with Excel.* New York, N.Y.: John Wiley and Sons.

Sawilowsky, S. S. (2002). Fermat, Schubert, Einstein, and Behrens-Fisher: The probable difference between two means when $\sigma_1^2 \neq \sigma_2^2$. *Journal of Modern Applied Statistical Methods, 1*(2), 461-472.

Schenker, N. & Gentleman, J. F. (2001). On judging the significance of differences by examining the overlap between confidence intervals. *The American Statistician, 55*(3), 182.

Smith, I. (2002). A set of statistical distribution functions in Excel VBA, (Version 1.0.24). Copyright © Ian Smith 2002-2003

Sprinthall, R. C. (2000). Basic statistical analysis (6th Ed.). Needham Heights, MA: Allyn and Bacon.

Steiger, J. H. & Fouladi, R. T. (1997). Noncentral interval estimation and the evaluation of statistical models. In *What if there were no significance tests?* Mulaik, H. & Steiger, L. E. M. New Jersey.

Triola, M. F. (2001). Elementary statistics (8th Ed.). Addison-Wesley Longham.

Welford, B. P. (1962). Note on a method for calculating corrected sums of squares and products. *Technometrics, 4*(3), 419.