

XVII. STATISTICAL TESTS, TESTS OF SIGNIFICANCE AND TESTS OF A HYPOTHESIS	2
A TEST OF SIGNIFICANCE	2
THE BASIS OF THESE TESTS	2
A STATEMENT OF THE HYPOTHESIS.....	3
THE VIEWS	3
THE PROBABILITIES	3
TESTING THE HYPOTHESIS, TESTS OF SIGNIFICANCE.....	5
P FUNCTIONS PROVIDED BY EXCEL	5
SUPPORTING FUNCTIONS PROVIDED BY EXCEL.....	5
TEST CALCULATIONS USING FUNCTIONS, CELL EQUATIONS AND MATHEMATICAL TOOLS	6
DATA ANALYSIS ROUTINES PROVIDED BY EXCEL	6
REPORTED PROBLEMS:.....	7
TESTS ON THE ACCURACIES OF FUNCTIONS AND DATA ANALYSIS ROUTINES.....	7
TEST DATA.....	8
VARIANCE TEST BASED FUNCTIONS AND ROUTINES	10
TEST ON CHITEST.....	10
TEST ON THE FUNCTION FTEST:	10
TEST ON THE DATA ANALYSIS F-TEST: TWO-SAMPLE FOR VARIANCES:.....	11
MEAN TEST BASED FUNCTIONS AND ROUTINES	12
TESTING THE DIFFERENCE BETWEEN MEAN VALUES	12
THE MISSING DATA PROBLEM	13
THE BEHRENS-FISHER PROBLEM.....	13
TEST ON PAIRED SAMPLES FOR MEANS IN EXCEL;.....	21
TEST ON TWO SAMPLES FOR MEANS, EQUAL VARIANCES.....	24
TESTS ON TWO SAMPLES FOR MEANS, UNEQUAL VARIANCES	25
DATA ANALYSIS: Z-TEST: TWO SAMPLE FOR MEANS:	27
TESTS OF SIGNIFICANCE ON CORRELATIONS.....	28
EXCEL FUNCTIONS	28
EXCEL FUNCTION FISHER:.....	28
EXCEL FUNCTION FISHERINV.....	29

APPROPRIATE TESTS ON CORRELATION COEFFICIENT VALUES.....	29
1. TESTING IF THE CORRELATION COEFFICIENT FROM A SAMPLE IS FROM A POPULATION THAT HAS A ZERO CORRELATION BETWEEN PAIRS.	29
2. TESTING IF THE CORRELATION COEFFICIENT FROM A SAMPLE IS FROM A POPULATION HAVING AN EXPECTED NON-ZERO CORRELATION COEFFICIENT BETWEEN PAIRS	30
3. TESTING IF THE CORRELATION COEFFICIENTS FROM TWO SEPARATE SAMPLES ARE FROM THE SAME POPULATION HAVING AN UNKNOWN CORRELATION COEFFICIENT BETWEEN PAIRS	30
4. COMBINING THE SAMPLE CORRELATION COEFFICIENTS FROM TWO SEPARATE INDEPENDENT SAMPLES HYPOTHESIZED AS BEING FROM THE SAME POPULATION, IN ORDER TO ESTIMATE THE POPULATION CORRELATION COEFFICIENT BETWEEN PAIRS	30

XVII. STATISTICAL TESTS, TESTS OF SIGNIFICANCE AND TESTS OF A HYPOTHESIS

A TEST OF SIGNIFICANCE

In any significance test the actual observed dataset is compared to all the possible datasets that might have been observed according to a specified stopping rule. (Bernard 1996). The null hypothesis then is a logical statement about a property of the sample with respect to the same property of a hypothesized population that is the source of all possible datasets. The null hypothesis then has an inherent aspect of being either true or false in the logical sense. That is, either the sample came from the hypothesized population or it did not. Usually both a null hypothesis and an alternate hypothesis is stated, so if the null hypothesis is true, the alternate is false or vice versa. This is not always the case. (see section 3 of Christensen 2005),

If the dataset is a random sample, then a probability distribution of the occurrence of the given dataset can be stated, with respect to all possible datasets. A decision about the hypotheses can then be based on this probability value.

If the dataset cannot be assumed to be random, then the test of significance is beyond the use of the standard functions in Excel.

THE BASIS OF THESE TESTS

Excel provides functions and routines that will perform a test of significance on a set of data. Note X gives information on what Excel can do for tests of significance found in several textbooks. Support by Excel is quite limited here in support of statistical courses and textbooks.

One of the central concepts here is that the test of significance is always about a specified hypothesis. This allows for making a decision about the hypothesis (as being true or false), based on an analysis of a data set, based on a specific test of significance.

The routines in the Data Analysis Tool-Pac return information about the data set including a p value, while the functions only return a p value. One cannot adequately interpret the p value in terms of a decision without having a hypothesis. The user has to make up the hypothesis and draw a conclusion based on the test results.

A STATEMENT OF THE HYPOTHESIS

The statistical tests in Excel, hinge on the basic, simpler concepts and practices related to the testing of a hypothesis. It cannot be said strongly enough, that we are looking at and using only a basic, elementary set of hypothesis statements and tools in Excel. There are faults and errors attributed to Excel here that upon investigation are found to be user faults with the hypothesis concept and with incorrect tests and with incorrect conclusions.

The hypothesis is a logical statement about some statistical measure of one or more populations. It is important to recognize that a hypothesis is a statement about some belief, and that the “data” provides “evidence” on how we determine the “validity” of the “belief”. The construct is binary (true-false) in the basic, elementary sets that can be tested for in Excel. The more involved, complicated relationships that are of interest in science and in medicine are beyond Excel to properly explore. Excel does not provide effect size value functions, but effect sizes can be constructed from cell equations. Excel does not have the non-central probability distributions that are needed when effect sizes are compared.

The written material below is very brief. One should read Hubbard and Bayarri (2003), Christensen (2005), and other sources.

Note Y is a guide to preparing a simple hypothesis that can be tested within Excel.

THE VIEWS

There are two views of a hypothesis and what it means. There is the Fisherian hypothesis and the Neyman-Pearson hypothesis. A mixing of the two is commonly found in practice, but the “combination” fails basic, fundamental statistical and logical concepts and is frequently misunderstood. The mixed version in practice creates more confusions than enlightenment. The Fisherian deals with the chances of a random occurrence as a p value, while the Neyman-Pearson deals with the error in making a decision to accept or reject the hypothesis”. A Fisherian hypothesis has only one term, “The Hypothesis”. The Neyman-Pearson hypothesis has two terms, a “Null Hypothesis” and an “Alternate Hypothesis”.

Another difference is that the Fisherian view is that if a significant chance occurred, then there would be new, additional research done (by the investigator or by others) to confirm the findings as being a valid scientific finding or just a chance occurrence. The reference is just one p value. A chance occurrence here can only be verified by additional research.

In the N-P view, additional research is not implied, and the conclusions about the hypothesis are taken as a “belief” of the investigator in the form of an acceptance of the null hypothesis or of the alternate hypothesis.

THE PROBABILITIES

Take U as a quantitative measure of the dataset. It could be any statistical measure.

Under Fisher's view, there was a hypothesized distribution (HD) that described the random distribution of U. There was a baseline value T which was the expected value of U

$$1) P = HD\{U | T\}$$

A decision was based on the value of P. The alpha value was just a decision point regarding the action to be taken.

Under the Neyman-Pearson test, there are two hypotheses, H0 and Ha. The test is on H0

$$2) P_0 = p\{U | H_0\} \text{ (The null)}$$

With respect to a preset alpha value. If P0 is less than alpha, the null is rejected and the alternate is accepted. Otherwise H0 is accepted.

The proper form of the decision would be:

$$3) P_0 = p\{H_0 | U\}$$

The only way to obtain 3) from 2) is to use a Bayesian method.

There are two decision probabilities here, an alpha and a beta.

Action	Probability That The			
	Null Is Actually		Alternate is Actually	
	True	False	True	False
Reject The Null	Alpha	1-Alpha	1-Alpha	Alpha
Accept The Null	1-Beta	Beta	Beta	1-Beta

The alpha and beta values are preset before the test is conducted, and n the size of the data set is also preset (n and beta are related).

Excel only provides the computations of a single p value, from the data. The user has to construct the hypothesis so that a reduction of the data to a p value can be used to accept or reject the null hypothesis. Note Y goes into detail on properly constructing a valid hypothesis, within the Excel environment.

Excel does not provide a Bayesian method for testing a hypothesis.

The accept/reject decision is based on comparing the P value from 2) to an alpha value (usually 0.05). All ensuing properties are based on the alpha value, not on the p value.

The P value here becomes vanishing small as a point estimate, which would be the case for equality between U and H. This is the inherent logical problem of tests for equality, which depend on sample size. The alternates \leq or \geq are often used to avoid the logical equality problem. These are one tail probabilities, whereas equal is a two tail probability. Some teachers advocate using the equality and then basing the p value on either one or two tailed probability values. The problem however is still there when testing for equality of two different treatments. In practice, the experiment testing for equality is designed (setting n) so that a two tailed test produces desired results.

See Christensen 2005 for an expanded discussion.

TESTING THE HYPOTHESIS, TESTS OF SIGNIFICANCE

We get a p value from a test of significance on the data. The null hypothesis as stated is then taken as the conditional basis for $p = p\{\text{data} \mid \text{hypothesis}\}$. A decision is then made to either accept the null hypothesis or reject it, based on the p value.

P FUNCTIONS PROVIDED BY EXCEL

CHITEST (Function) –

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.

CHITEST is a rather limited function. Microsoft could have done a better job here of allowing a one dimensional Goodness-of-fit test, and having the function calculate the independent expected values directly from the observed value contingency table. Some of the Add-ins fill this gap.

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

TTEST (Function) - 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

ZTEST (Function) - 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.

SUPPORTING FUNCTIONS PROVIDED BY EXCEL

These are some of the other functions that can be used in worksheet cells to provide values as part of the data analysis and significance testing process.

AVERAGE – Returns the average of a list of data points (or arguments), numbers only.

COUNT – Counts how many numbers there are in a range.

COUNTIF – The COUNT function where a cell in the range input has to meet a single criteria (number, logical expression or text) in order to be counted.

STDEV – Returns the standard deviation estimate of a population based on a sample data set of numbers. Divides by n-1.

NORMDIST – Returns the normal distribution p value given values for μ and σ .

NORMSDIST – Returns the normal distribution p value for a given z value.

TDIST – Returns the t distribution p value given

FDIST – Returns the F distribution p value given

CHIDIST – Returns the Chi Square distribution p value

CORREL - Returns the correlation coefficient between two separate ranges of paired data.

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

FISHER – Transforms a correlation (+ or -) r value to a z value, using Fishers transform.

FISHERINV – The inverse of Fisher’s function.

CONFIDENCE –

Calculates a half width interval about the sample mean using the population standard deviation and a z value.

If the population standard deviation is not known, then the Data Analysis routine “Descriptive Statistics” should be used with the “range” of the dataset entered and the “Confidence Interval for Mean” checked with the desired percentile in the right-hand box. The Data Analysis Confidence Level is a calculation of an interval about the sample mean using the sample standard deviation and a t value from the input p value. The default interval is one that is likely to contain the population mean at an alpha level.

TEST CALCULATIONS USING FUNCTIONS, CELL EQUATIONS AND MATHEMATICAL TOOLS

Note Z shows in detail how simple cell tables can be constructed to arrive at a correct p value, using the above functions.

DATA ANALYSIS ROUTINES PROVIDED BY EXCEL

F-TEST TWO-SAMPLE FOR VARIANCES:

Input Variable 1 Range: A block defining the input data. A list of values.

Input Variable 2 Range: A block defining the input data. A list of values.

Labels: Does the first cell in the list have a label?

Alpha:

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

T-TEST: PAIRED TWO SAMPLE FOR MEANS:

Input Variable 1 Range: A block defining the input data. A list of values.

Input Variable 2 Range: A block defining the input data. A list of values.

Hypothesized Mean Difference: A number

Labels: Does the first cell in both lists have a label?

Alpha: The probability of rejection of the hypothesis.

T-TEST: TWO SAMPLE ASSUMING EQUAL VARIANCES:

Input Variable 1 Range: A block defining the input data. A list of values.

Input Variable 2 Range: A block defining the input data. A list of values.

Hypothesized Mean Difference: A number

Labels: Does the first cell in both lists have a label?

Alpha: The probability of rejection of the hypothesis.

T-TEST: TWO-SAMPLE ASSUMING UNEQUAL VARIANCES:

Input Variable 1 Range: A block defining the input data. A list of values.

Input Variable 2 Range: A block defining the input data. A list of values.

Hypothesized Mean Difference: A number

Labels: Does the first cell in both lists have a label?

Alpha: The probability of rejection of the hypothesis.

Z-TEST: TWO SAMPLE FOR MEANS:

Input Variable 1 Range: A block defining the input data. A list of values.

Input Variable 2 Range: A block defining the input data. A list of values.

Hypothesized Mean Difference: A number

Variable 1 Variance (known):

Variable 2 Variance (known):

Labels: Does the first cell in both lists have a label?

Alpha: The probability of rejection of the hypothesis.

REPORTED PROBLEMS:

There was little in the literature regarding statistical tests using EXCEL. Table 17-1 was all that was found.

Table 17-1: Excel Problems About Statistical Tests

Application or Function	Problem	Source	Fix or Comments
Paired t Test, with missing data	Incorrect values	Simon 2000	Do not have any missing data in any range specified as an input to a statistical computation. It is very important to recognize that the paired t test can only be applied to a set of data pairs. The test is not valid when missing data appears.
Two sample t test, unequal variances	Df's are rounded not truncated.	RSS 1996	Test is an approximation. See discussion below on the Fisher-Behrens problem
Confidence Function	Uses z instead of t	RSS 1996	The CONFIDENCE function is only to be used with a known normal population standard deviation. Use of z is correct.

TESTS ON THE ACCURACIES OF FUNCTIONS AND DATA ANALYSIS ROUTINES

There are differences between the accuracies 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 lot of space to show both 2000 and 2003 outputs, only the Excel 2003 values are shown. They are the same for Excel 2007.

The test results on Excel 2003 apply to Excel 2007. There were no changes made to these functions/routines for Excel 2007.

TEST DATA

Table 17-2: Data Sets Used For Tests

Set 1		Set 2		Set 3		Set 4	
A	B	C	D	E	F	G	H
3	2	39	43	a+0.4792	a+0.4130	49.17	53.38
4		49	52	a+0.4562	a+0.3635	51.12	60.72
3	2	49	54	a+0.4217	a+0.8865	53.45	99.93
	3	51	60	a+0.1971	a+1.0466	45.56	109.78
2	3	52	61	a+0.9054	a+0.3824	38.49	4.92
4	3	53	61	a+0.1285	a+0.4194	60.20	32.85
4	3	53	64	a+0.7258	a+0.8125	40.69	92.77
3	4	54	67	a+0.1267	a+0.8649	67.62	121.89
2	3	58	67	a+0.3427	a+0.6979	44.88	48.39
4	2	58	67	a+0.1079	a+1.1558	85.20	24.00
4	2	59	67	a+0.4181	a+0.5900	53.07	78.60
3	2	60	67	a+0.1848	a+0.8718	56.64	28.96
4	3	60	68	a+0.0424	a+1.0104	46.23	26.75
3	2	61	69	a+0.9731	a+0.2760	53.04	1.69
2	2	62	69	a+0.4839	a+0.7267	58.30	36.77
3	2	62	70	a+0.9565	a+0.2598	43.24	97.40
3	4	64	71	a+0.7153	a+1.0490	47.12	124.11
4	2	66	71	a+0.5156	a+0.2710	38.19	111.61
4	2	67	72	a+0.6408	a+0.3856	44.32	84.47
2	3	End	73	a+0.7284	a+0.5472	38.19	62.49
End	End		73	a+0.3760	a+1.1153	48.14	End
			74	a+0.9309	a+1.0629	60.46	
			79	a+0.0894	a+0.6591	62.75	
			82	a+0.1265	a+0.7823	67.10	
			84	a+0.8862	a+0.5469	42.65	
			End	a+0.2160	a+0.2799	42.73	
				a+0.6907	a+0.4830	47.96	
				a+0.7906	a+1.1009	51.99	
				a+0.5438	a+0.7111	55.24	
				a+0.9128	a+1.1304	36.91	
				End	End	End	

Set 1 is a set of paired data with missing values. The data in the columns is not normal

Set 2 represents two unequal data sets. The means, variances and numbers of observation for C and D are all different, but the differences are not large. The data in the columns is not normal

Set 3 comes from one population with equal means and equal variances (however the sample means and standard deviations are different). The ‘a’ value (i.e. additive) is normally 1000. The data in the columns is not normal

Set 4 comes from two different populations, with equal means but different variances. The data in the columns is normal.

Set 5 is a variable length set with two columns. The first (Column I) represents the control data set, and the second (Column J) represents the treatment data set. The base case is where the numbers in both columns are all random normal (0,1) z values from an accurate normally distributed random source¹. Whole numbers can be added to the z values in the manner of the NIST StRD, SmLs01 to SmLs09 series of test data sets. In addition, the differences between the columns can be set to achieve a specified effect size (Hedges g value, Kline 2004, page 101). Set 5 exists as an Excel worksheet; too large to be replicated here, since normally the length (n) is 2000.

Set 6 (table 3-2) is the counts of responses to political efficacy questions from Jöreskog (2002). The data comes from Barnes and Kaase (1979), and is the USA data and questions.

Table 3-3: Ordinal Data As A Contingency Table For Tests, Set 6

Political Efficacy	AS	A	D	DS	DK	NA
NOSAY	175	518	857	130	29	10
VOTING	283	710	609	80	26	11
COMPLEX	343	969	323	63	9	12
NOCARE	250	701	674	57	20	17
TOUCH	273	881	462	26	60	17
INTEREST	264	762	581	31	62	19

NOSAY: People like me have no say in what the government does.

VOTING: Voting is the only way that people like me can have any say about how the government runs things.

COMPLEX: Sometimes politics and government seem so complicated that a person like me cannot really understand what is going on.

NOCARE: I don’t think that public officials care much about what people like me think.

TOUCH: Generally speaking, those we elect to Congress in Washington lose touch with the people pretty quickly.

INTEREST: Parties are only interested in people’s votes but not in their opinions.

The responses allowed were:

AS Agree strongly

¹ Marsaglia’s MWC256 random number generator, (Marsaglia 1995 and 2002) coupled with Smith’s precise inverse normal function. (Smith 2002)

- A Agree
- D Disagree
- DS Disagree strongly
- DK Don't know
- NA No answer

VARIANCE TEST BASED FUNCTIONS AND ROUTINES

TEST ON CHITEST

To use the CHITEST function the first step is to put the data into a range of Excel cells. The second step is to build a corresponding (new) table giving the expected values for each of the cells in the data range. The user has to go to a statistics book to find out how the expected values table is built from the data table. Both tables must express frequency as counts. Relative frequencies will give incorrect outputs.

Table 3-4: Expected Values For Data Set 6

Political Efficacy	AS	A	D	DS	DK	NA
NOSAY	264.667	756.833	584.333	64.500	34.333	14.333
VOTING	264.667	756.833	584.333	64.500	34.333	14.333
COMPLEX	264.667	756.833	584.333	64.500	34.333	14.333
NOCARE	264.667	756.833	584.333	64.500	34.333	14.333
TOUCH	264.667	756.833	584.333	64.500	34.333	14.333
INTEREST	264.667	756.833	584.333	64.500	34.333	14.333

The third step is to enter =CHITEST(data table range, expected data range) into another cell. For the above data, the result is a p value of 4.00374523030632E-129. Doing the calculations defined as the Chi-Square Goodness Of Fit Test and using the Excel function CHIDIST, a value of 4.00374523030632E-129 resulted. If the reference chi square function is used instead, a value of 4.003745229443920E-129 is obtained. The difference reflects the error in CHIDIST and is expressed as an LRE value of 9.67.

The conclusion is that the Excel algorithm in the CHITEST function is the correct one.

Errors then are from errors in the expected values table and in the CHIDIST function. CHIDIST function errors are covered in section 12-3.

CHITEST returns correct values if the Expected Values table is correct.

TEST ON THE FUNCTION FTEST:

The Microsoft KBAs indicate that the FTEST function just computes the ratio of two variances where the variances come from the VAR function and then using the FDIST function to obtain a p value. The VAR function holds up well against overload as shown below.

Table17-5: Excel 2003 LRE Values For VAR. Deviations N(0.2,0.1) Distributed.

Additive	0	1	10	100	1000	10000	100000	1000000	10000000
VAR	15.00	12.93	14.08	14.14	14.48	12.35	11.54	10.92	9.69

Section 5 gives more information on the accuracies of the VAR function.

Given the ratio, the FINV function then is 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

The FTEST function was “built” as only testing for equality of two variances. Consequently the function returns a p value on the equality of the variances in the form of a two-tail test. One minus the returned p value indicates the probability of being unequal.

For the question of one variance being less than the other, a one tailed test has to be done. Here the appropriate p value is on-half the returned two-tailed p value.

Table 17-6: FTEST Function Response

Cell Entry	Returned Value	One-Tail p Value
=FTEST(C,D)	0.9425381810184540	0.481410961628470

The internal FINV function returns correct values as shown in the following table.

Table 17-7: Table F Distribution Critical Values

Source	Abramson 1963	Larson 2003	Levine 1999	Pelosi 2000	Moore 2003	Reference F Values	Excel FINV Values	Reference Function q values
Stated Tail	Q(F v1, v2)	Upper tail area	Upper tail area	Area in the right tail	P lying to the right			
Alpha	0.05	0.05	0.05	0.05	0.05			
6 – 6	4.28	4.28	4.28	4.284	4.28	4.28	0.0500979204874610	0.050097920484424
12 – 20	2.28	2.28	2.28	2.278	2.28	2.28	0.0497858488792365	0.049785848818623
20 – 10	2.77	2.77	2.77	2.774	2.77	2.77	.0502301740345887	0.050230174024306
60 – 120	1.43	1.43	1.43	1.429		1.43	0.0496728697381574	0.049672869608066

The table 17-7 F distribution values are all essentially the same, differing due to the limitations on textbook F ratios to only 3 digits. This table defines the accepted response from a correct F test. The q value is that tail area to the right of the F value.

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. There is then this ambiguity of what is the correct F test. The resolution is of course on how you are viewing the tails of the F distribution.

TEST ON THE DATA ANALYSIS F-TEST: TWO-SAMPLE FOR VARIANCES:

Table 17-8: 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	

Here Excel returns an accurate value. The True 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 FDIST.

MEAN TEST BASED FUNCTIONS AND ROUTINES

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 unknown 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) and three Data Analysis routines for statistical solutions for these three situations. The questions here are just what do these do, and do they compute the statistics correctly in terms of theory, and are the results numerically accurate. Other concerns are; how robust are they on non-normal data, how stable are the results in terms of type I error rates and what is the power.

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 loses 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 MISSING DATA PROBLEM

The dependent, paired values t test is the only one affected by missing data. Missing data essentially voids a paired comparison. There are different ways to deal with missing data, including artificially generating data values by imputation. Excel does not have any external imputation routines. Therefore when a blank cell or a cell having non-numeric data occurs, how Excel deals with it is important. The accepted method is to exclude any unpaired data, and only do the analysis on the remaining paired data. However Excel may not do it this way, so it is of interest on how Excel handles the problem and how accurate are the returned p values.

THE BEHRENS-FISHER PROBLEM

THEORETICAL SOLUTIONS

The theoretical problem is whether two entirely different populations having separate (unknown) means and variances have a difference in their means that is “statistically significant”. The concept of “statistical significance” depends on being able to characterize the difference in terms of a “difference population” which does not exist. Hence, theoretically this “difference population” is a fiction, an artificial creation. Only when the variances are equal (but unknown), can a theoretical solution be obtained (See Hogg and Craig 1978, sections 6.4 and 8.3).

Solutions based on distribution free tests (such as the Mann-Whitney-Wilcoxon test) do however have theoretical solutions (Hogg and Craig 1978, Chapter 9). Sawilowsky (2002), says, “I would be remiss if I failed to note that numerous Monte Carlo studies have shown that the nonparametric Wilcoxon Rank Sum test (the Mann-Whitney-Wilcoxon 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.” Excel does not provide this test, which can be considered a fault in Excel.

There are several approximations found in textbooks and in the literature for this problem, 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 $n(=N-1)$, 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 the on the null hypothesis considered, namely that the two populations means are equal, we have

$$x_1 - x_2 = (s_1 * t_1 / \sqrt{n_1}) - (s_2 * t_2 / \sqrt{n_2}) \quad (\text{Fisher's equation 53})$$

Where t_1 and t_2 are distributed independently in the two distributions.” (Fisher 1973b, p 98)

“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 1973B, 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. Both Fisher's and Schenker and Gentleman's methods fail to give acceptable conclusions.

THE WELCH-ASPIN-SATTERTHWAITE SOLUTION

The Welch-Aspin-Satterthwaite solution is a solution to the Behrens-Fisher problem. It is essentially the construction of a hypothetical new population in which both x_1 and x_2 are possible means. 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. Welch took Satterthwaite's method of forming a consolidated new variance distribution. It is commonly referred to as the Aspin-Welch test or the Welch test in research papers. However some statistics textbooks (i.e. McCabe and Moore 2003) will ignore all this and just use the term “pooled df ” for this test, or “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.

Balkin and Mallows (2001) use Johnson's asymmetrical t distribution on samples having large differences in sample sizes and asymmetry, and get results comparable to permutation-based tests. Departures from normality may not be that big a problem.

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.

Table 17-9 gives the common four methods for computing a df value.

Table 17-9: 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

Methods to obtain a t value from the difference between the two means are listed in table 17-10. There are others such as the Score statistic that are not considered here, since they are not found in or introduced in introductory statistics textbooks.

Table 17-10: 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	$\text{var}_{\text{pooled}} / n_1$	$\text{var}_{\text{pooled}} / n_2$	Equal variances, pooled t test
3	$\text{var}_1 / (2n_1 + 1)$	$\text{var}_2 / (2n_2 + 1)$	Fisher's 1939 form
4	$\text{var}_1 * (n_1 - 1) / (n_1^2 - 3n_1)$	$\text{var}_2 * (n_2 - 1) / (n_2^2 - 3n_2)$	Fenstad's Statistic
5	$\text{var}_1 * (n_1 - 1) / n_1^2$	$\text{var}_2 * (n_2 - 1) / n_2^2$	Wald Statistic

Where:

var_1 = Variance of Sample 1

var_2 = Variance of Sample 2

$\text{var}_{\text{pooled}} = ((n_1 - 1) * \text{var}_1 + (n_2 - 1) * \text{var}_2) / (n_1 + n_2 - 2)$

The t value = Difference in Means / Square Root (Term 1 + Term 2) [Methods 1-4]

The t value = (Differences in Means)² / (Term 1 + Term 2) [Method 5]

Currently, only t-value methods 1 and 2 are considered. This then gives six ways to calculate a p value

Table 17-11: 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 is at sample sizes related to the ratio of the known variances of the samples.

$$\kappa = \text{variance population 2} / \text{variance population 1}$$

$$n_1 / (n_1 + n_2) = 1 / (1 + \sqrt{\kappa})$$

However the local optimal design is sensitive to the misspecification of the κ value.

THE WELCH DF VALUE

Welch’s df (df method 4):

$$\text{Let } u_1 = (s_1 * s_1) / n_1 \text{ and } u_2 = (s_2 * s_2) / n_2$$
$$\text{df} = (u_1 + u_2)^2 / [(u_1^2 / (n_1 - 1)) + (u_2^2 / (n_2 - 1))]$$

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

$$\text{Let } u_1 = (s_1 * s_1) / n_1 \text{ and } u_2 = (s_2 * s_2) / n_2$$

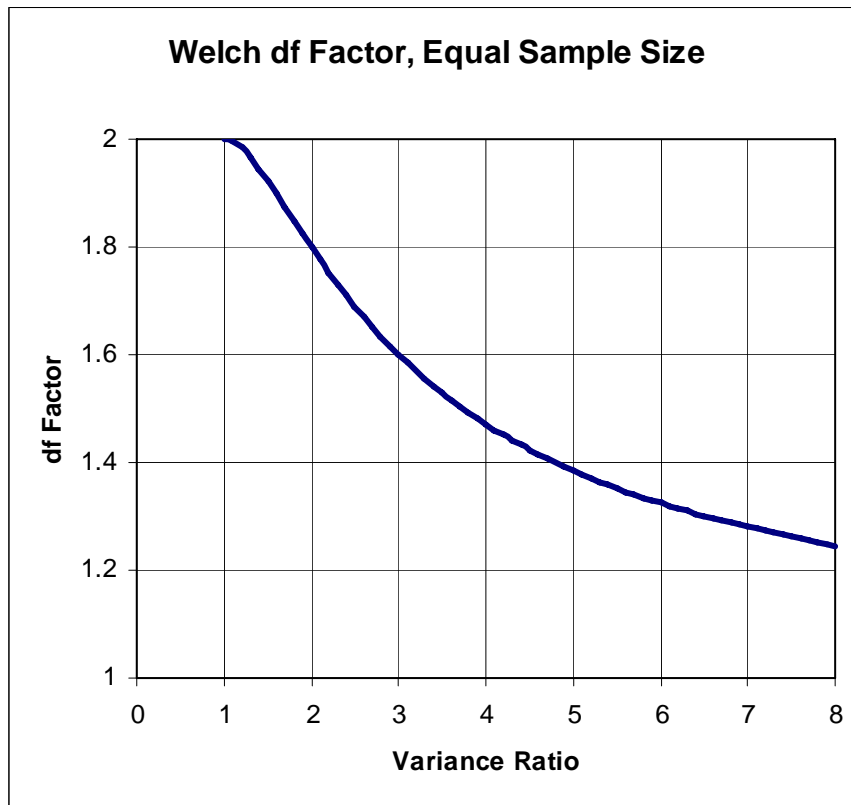
$$C = u_1 / (u_1 + u_2)$$

$$\text{Let } tm_1 = c^2 / (n_1 - 1) \text{ and } tm_2 = (1-c)^2 / (n_2 - 1)$$

$$\text{Then Aspin-Welch eta} = 1 / (tm_1 + tm_2) \quad [\text{Zukov (nd)}]$$

Figure 17-1 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. n-1) gives the Welch df value.

Figure 17-1: Welch df Factor



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 test² 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 we do not truly know that the variances are equal.

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. McCabe and Moore (2003) recommend 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

² There have been many papers over the years that point out that if you use the F test to decide on equal/unequal variances at an alpha level, and then do the t test at the same alpha level, you have totally lost control of the type I and type II error rates. The classical adjustments for the type I error rate here; do not hold up on simulations, primarily because of the sensitivity of the F test to even small deviations from normality. See Sawilowsky 2002.

fractional df values. Both Excel's and Smith's t distribution functions will only allow integer df values to be entered.

One of the criticisms of Excel in the literature (RSS 1996) was that Excel used an integer df value rather than the calculated Welch df value. Actually in the Excel TTEST function (under option 3) the fractional Welch df value is used (see table 17-27) and in the Data Analysis Tool Pac T Test, the df value is rounded to an integer (see tables 17-28, 17-29 and 17-30). Consequently the two methods will not arrive at the same P value for the identical data sets.

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 generally follows df method 4.

THE T STATISTIC MEASURES

Best and Rayner (1987) identify four t statistic measures that can be used:

(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) define the other three (W, L and S), but concluded that for their $n_1=4$ and $n_2=8$ sample sets (from Monte Carlo sets), the power of the test for differences was about the same.

Best and Rayner (1987) define calculation method 3 as the V statistic. They find 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 it can be used for both tests involving equal and unequal variances.

SOME TEXTBOOK DIRECTIONS

McCabe and Moore (2003) say use calculation 4 for equal variances and calculation 2 or 3 for unequal variances. Calculation 3 is preferred for unequal variances.

Larson and Farber (2003) say use calculation 4 for equal variances and calculation 2 for unequal variances

Triola (2001) says use calculation 4 for equal variances and calculation 2 for unequal variances

Lind, Marchal and Mason (2001) say use calculation 4 for equal variance (does not cover unequal variances)

Pelosi and Sandiffer ((2000), say use calculation 4 for equal variances and calculation 3 for unequal variances.

Levine, Berenson and Stephen (1999) say use calculation 4 for equal variances. Unequal variances are not covered.

The general consensus among textbooks is use calculation 4 for equal variances, since 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 tables 17-13, 17-15 and 17-17, calculation method 2 gives higher p values than calculation method 3. Consequently using textbook recommendations may not be the best solution method.

THE BEST APPROACH

In applied studies and research, the current view is that the real problem is where 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. 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 three reference data sets are given in tables 17-12 thru 17-17. These are standard sets of values to compare with Excel test results. The Excel TDIST function is used in all cases, to avoid any test errors due to the inherent TDIST function errors. Excel function and routine outputs that differ from these values then are errors in the algorithms.

Table 17-12: Reference Values for Test Data Set 2

Measure	Col C	Col D	Summary	Values
Count	19	25	Sum	44
Mean	56.68421	67.40000	Difference	1.071578947368420E+01
Variance	48.45029	81.25000	Pooled	6.719298245614030E+01
			Denom, Method 1	2.408322110796350E+00
			Denom, Method 2	2.494833072390840E+00

Table 17-13: Calculated Values for Test Data Set 2

Calculation Method	df	T Value	df Value	t Value	tDIST one tail	TDIST Two Tail
1	2	1	42	4.449483491285	3.10967714824979E-05	6.21935429649958E-05
2	3	1	18	4.449483491285	1.54836265793784E-04	3.09672531587569E-04
3	4	1	41.97893625	4.449483491285	3.11216827970778E-05	6.22433655941556E-05
4	2	2	42	4.295192969931	5.05030693886014E-05	1.01006138777203E-04
5	3	2	18	4.295192969931	2.17858154882741E-04	4.35716309765482E-04
6	4	2	41.97893625	4.295192969931	5.05394581922119E-05	1.01078916384424E-04

Table 17-14: Reference Values for Test Data Set 3

Measure	Col E	Col F	Summary	Values
Count	30	30	Sum	60
Mean	1000.50377	1000.69673	Difference	1.92960000003120E-01
Variance	0.09206	0.09046	Pooled	9.125842216092200E-02
			Denom, Method 1	7.799932570688120E-02
			Denom, Method 2	7.799932570688120E-02

Table 17-15: Calculated Values for Test Data Set 3

Calculation Method	df	T Value	df Value	t Value	TDIST one tail	TDIST Two Tail
1	2	1	58	2.473867539900	8.15718183270718E-03	1.63143636654144E-02
2	3	1	29	2.473867539900	9.73156714048129E-03	1.94631342809626E-02
3	4	1	57.99557946	2.473867539900	8.15730026652737E-03	1.63146005330547E-02
4	2	2	58	2.473867539900	8.15718183270718E-03	1.63143636654144E-02
5	3	2	29	2.473867539900	9.73156714048129E-03	1.94631342809626E-02
6	4	2	57.99557946	2.473867539900	8.15730026652737E-03	1.63146005330547E-02

Table 17-16: Reference Values for Test Data Set 4

Measure	Col G	Col H	Summary	Values
Count	30	20	Sum	50
Mean	51.02167	65.07400	Difference	1.40523333333330E+01
Variance	113.58181	1517.89320	Pooled	6.694550686805560E+02
			Denom, Method 1	8.926405797383880E+00
			Denom, Method 2	7.469131300897470E+00

Table 17-17: Calculated Values for Test Data Set 4

Calculation Method	df	T Value	df Value	t Value	TDIST one tail	TDIST Two Tail
1	2	1	48	1.574243167104	6.10000865396000E-02	1.22000173079200E-01
2	3	1	29	1.574243167104	6.31395491587430E-02	1.26279098317486E-01
3	4	1	20.90884984	1.574243167104	6.52224127169738E-02	1.30444825433948E-01
*			20	1.574243167104	6.55589794414705E-02	1.31117958882941E-01
*			21	1.574243167104	6.51886578433326E-02	1.30377315686665E-01
4	2	2	48	1.881387910753	3.29957252874018E-02	6.59914505748037E-02
5	3	2	29	1.881387910753	3.49963541495991E-02	6.99927082991982E-02
6	4	2	20.90884984	1.881387910753	3.69599645055018E-02	7.39199290110035E-02

*P values with the df truncated and rounded to integers

TEST ON PAIRED SAMPLES FOR MEANS IN EXCEL;

There are two issues here, accuracy regarding a data set without any missing values, and a second, on how Excel deals with missing values. The second issue involves an issue on how to act on missing data, and the resulting accuracy based on normal conventions regarding missing data.

Users have reported errors here (Simon 2000 and Simonoff 2000). The theory on t tests on paired data does not allow for unpaired data, such as that in the second and fourth rows of data set 1. Other software programs such as MINITAB are able to handle missing data. The accepted method here is to disregard any rows having missing data, and only do calculations on paired data.

In data set 1, there is missing data in rows 2 and 4. Data set 1 with rows 2 and 4 removed will be referred to as Data Set 1-mod.

Another problem that surfaces is how does one represent missing data in an Excel cell. Both Simon (2000) and Simonoff (2000) represent it by an asterisk in the column. In Excel when an asterisk or space is encountered, it is interpreted as a text or non-numeric value. There are other users who represent it by a “blank” cell. Here Excel interprets a blank cell as an “empty” cell.

Case 1 is when the data is entered cell by cell in a column and an asterisk is entered for missing data (new entry).

Case 2 is when the data is entered cell by cell in a column and a cell is skipped if the data is missing (new entry).

Case 3 is when case 1 entry is completed and computed, and then asterisks are changed to empty cells (using the delete key).

Case 4 is when case 2 entry is completed and computed, and then blank cells have an asterisk entered.

**THE TTEST FUNCTION WITH TYPE OPTION 1 (PAIRED)
NO MISSING DATA VALUES**

Table 17-18: TTEST Function Results, Option 1, Set 3

Cell Entry	Returned Value	LRE On True Value
=TTEST(E,F,1,1)	0.0158576391948868	7.86
=TTEST(E,F,2,1)	0.0317152782097736	7.86

TTEST returns correct values.

WITH MISSING DATA VALUES

Table 17-19: TTEST Function Results, Option 1, Set 1, Cases 1, 2, 3 and 4

Cell Entry	Returned Value	LRE On True Value
=TTEST(A,B,1,1)	0.01843903351664	7.91
=TTEST(A,B,2,1)	0.03687806703328	7.91

TTEST returns correct values.

**DATA ANALYSIS: T-TEST: PAIRED TWO SAMPLE FOR MEANS:
NO MISSING DATA VALUES**

Table 17-20: T Test: Paired Two Sample For Mean, Actual Excel Output

	E	F
Mean	1000.503767	1000.696727
Variance	0.092055155	0.090461689
Observations	30	30
Pearson Correlation	-0.201427066	
Hypothesized Mean Difference	0	
Df	29	
t Stat	-2.256987345	
P(T<=t) one-tail	0.015857639	
t Critical one-tail	1.699126996	
P(T<=t) two-tail*	0.031715278	
t Critical two-tail	2.045229611	

*Note that 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 $T \neq t$.

Data Analysis t-Test: Paired Two Sample for Means: Returns correct values if there is NO MISSING DATA

WITH MISSING DATA VALUES

Set 1, Case 1 or Case 4:

The error message “t Test: Paired Two Sample for Means – Input Range Contains Non-numeric Data” appears in a message box. The erroneous output given by Simonoff (2000) on page 6 does not appear. Deleting the non-numeric asterick (or space bar), allowed a rerun to give results rather than an error message.

Table 17-21: T Test: Paired Two Sample For Mean, Actual Excel Output, Set 1, Case 2 and 3

t-Test: Paired Two Sample for Means		
	A	B
Mean	3.210526316	2.578947368
Variance	0.619883041	0.479532164
Observations	19	19
Pearson Correlation	-0.176998081	
Hypothesized Mean Difference	0	
Df	18	
t Stat	1.714285714	
P(T<=t) one-tail	0.051821506	
t Critical one-tail	1.734063592	
P(T<=t) two-tail	0.103643013	
t Critical two-tail	2.100922037	

Table 17-22: p Value errors from Table 17-21

Cell Entry	True Value	LRE On True Value
0.051821506	0.01843903351664	0
0.103643013	0.03687806703328	0

This is the same table as shown in Simonoff (2000) on page 7. If the missing data rows are removed, table 17-21 occurs which has the correct values

Table 17-23: T Test: Paired Two Sample For Mean, Actual Excel Output, Set 1-mod

t-Test: Paired Two Sample for Means		
	Sample 1	Sample 2
Mean	3.166666667	2.555555556
Variance	0.617647059	0.496732026
Observations	18	18
Pearson Correlation	-0.176998081	
Hypothesized Mean Difference	0	
df	17	
t Stat	2.264878763	
P(T<=t) one-tail	0.018439034	
t Critical one-tail	1.739606716	
P(T<=t) two-tail	0.036878067	
t Critical two-tail	2.109815559	

In comparing table 17-19 with table 17-21, one can see that the Data Analysis routine incorrectly handles cells with missing data.

Data Analysis t-Test: Paired Two Sample for Means: Returns incorrect values when there is missing data.

TEST ON TWO SAMPLES FOR MEANS, EQUAL VARIANCES

THE TTEST FUNCTION WITH TYPE OPTION 2 (EQUAL VARIANCES)

Method Used in Excel: Calculation 4; (df = $n_1 + n_2 - 2$, var = $((n_1 - 1) * \text{var}_1 + (n_2 - 1) * \text{var}_2) / (n_1 + n_2 - 2)$)

Table 17-24: TTEST Function Results, Option 2, Equal Variances

Cell Entry	Returned Value	LRE On True Value
=TTEST(C,D,1,2)	5.05030693886014E-05	16
=TTEST(C,D,2,2)	1.01006138777203E-04	16
=TTEST(E,F,1,2)	8.15718183270718E-03	16
=TTEST(E,F,2,2)	1.63143636654144E-02	16
=TTEST(G,H,1,2)	3.29957252874018E-02	14.7
=TTEST(G,H,2,2)	6.59914505748036E-02	14.7

TTEST returns correct values

DATA ANALYSIS: T-TEST: TWO SAMPLE ASSUMING EQUAL VARIANCES:

Method Used in Excel: Calculation 4; {df= $n_1 + n_2 - 2$, var = $((n_1 - 1) * \text{var}_1 + (n_2 - 1) * \text{var}_2) / (n_1 + n_2 - 2)$ }

Table 17-25: T-Test: Two Sample Assuming Equal Variances Test Results, Actual Excel Output

	E	F
Mean	1000.503767	1000.696727
Variance	0.092055155	0.090461689
Observations	30	30
Pooled Variance	0.091258422	
Hypothesized Mean Difference	0	
Df	58	
t Stat	-2.47386754	
P(T<=t) one-tail	0.008157182	
t Critical one-tail	1.671552763	
P(T<=t) two-tail*	0.016314364	
t Critical two-tail	2.001717468	

LRE on P(T<=t) one-tail, 16.0

Table 17-26: T-Test: Two Sample Assuming Equal Variances Test Results, Actual Excel Output

	G	H
Mean	51.02166667	65.074
Variance	113.5818075	1517.893204
Observations	30	20
Pooled Variance	669.4550687	
Hypothesized Mean Difference	0	
Df	48	
t Stat	-1.881387911	
P(T<=t) one-tail	0.032995725	
t Critical one-tail	1.677224197	
P(T<=t) two-tail*	0.065991451	
t Critical two-tail	2.010634722	

LRE on P(T<=t) one-tail, 16.0

Data Analysis t-Test: Two Sample Assuming Equal Variances: Returns correct values.

TESTS ON TWO SAMPLES FOR MEANS, UNEQUAL VARIANCES

THE TTEST FUNCTION WITH TYPE OPTION 3 (UNEQUAL VARIANCES)

P Values, TTEST Excel 2003

Table 17-27: TTEST Function Results, Option 3, Equal Variances

Cell Entry	Returned Value	LRE On TDIST Value	LRE On True Value
=TTEST(C,D,1,3)	3.112078236123900E-05	4.54	9.00
=TTEST(C,D,2,3)	6.224156472247800E-05	4.54	9.00
=TTEST(E,F,1,3)	0.008157298166861160	6.59	7.68
=TTEST(E,F,2,3)	0.016314596333722300	6.59	7.68
=TTEST(G,H,1,3)	6.52209588667085E-02	4.65	9.34
=TTEST(G,H,2,3)	1.30441917733417E-01	4.65	9.34

The results are interesting in that TTEST does not use the Excel TDIST function, but uses the Excel BETADIST function to calculate p values. T distribution values can be obtained from the incomplete beta where fractional degree of freedom values can be entered. The LRE values in column 4 are based on Smith's incomplete beta distribution, and they reflect the error in Excel's BETADIST function.

The conclusion here is that TTEST with option 3 correctly calculates the Welch-Aspin-Satterthwaite solution to the Fisher-Behrens problem of differences in means with unequal variances.

DATA ANALYSIS: T-TEST: TWO-SAMPLE ASSUMING UNEQUAL VARIANCES:

Table 17-28: Excel Output For Data Set 2, Actual Excel Output

	Variable 1	Variable 2
Mean	56.68421053	67.4
Variance	48.4502924	81.25
Observations	19	25
Hypothesized Mean Difference	0	
Df	42	
t Stat	-4.449483491	
P(T<=t) one-tail	3.10968E-05	
t Critical one-tail	1.681952358	
P(T<=t) two-tail*	6.21935E-05	
t Critical two-tail	2.018081679	

Table 17-29: Excel Output For Data Set 3, Actual Excel Output

	Variable 1	Variable 2
Mean	1000.503767	1000.696727
Variance	0.092055155	0.090461689
Observations	30	30
Hypothesized Mean Difference	0	
Df	58	
t Stat	-2.47386754	
P(T<=t) one-tail	0.008157182	
t Critical one-tail	1.671552763	
P(T<=t) two-tail*	0.016314364	
t Critical two-tail	2.001717468	

Table 17-30: Excel Output For Data Set 4, Actual Excel Output

	Variable 1	Variable 2
Mean	51.02166667	65.074
Variance	113.5818075	1517.893204
Observations	30	20
Hypothesized Mean Difference	0	
Df	21	
t Stat	-1.574243167	
P(T<=t) one-tail	0.065188658	
t Critical one-tail	1.720742871	
P(T<=t) two-tail*	0.130377316	
t Critical two-tail	2.079613837	

Table 17-31: Summary, t-Test: Two-Sample Assuming Unequal Variances

Data	True, unequal variances, Welch df	Returned one tailed P Value	LRE
Set 2	3.11216827970778E-05	3.10967714824979E-05	3.10
Set 3	8.15730026652737E-03	8.15718183270718E-03	4.84
Set 4	6.52224127169738E-02	6.51886578433326E-02	3.29

The Data Analysis t-test for unequal variances calculates the Welch df value, then rounds it to an integer and then uses the Excel TDIST function to obtain a p value. The LRE value reflects the error from rounding the Welch df value.

TTEST will give approximately correct values or incorrect values depending on what textbook you are using as the criteria. The key is the df value in the output table.

DATA ANALYSIS: Z-TEST: TWO SAMPLE FOR MEANS:

Table 17-32: z-Test Output For Data Set 2, Actual Excel Output

	E	F
Mean	1000.503767	1000.696727
Known Variance	0.3	0.3
Observations	30	30
Hypothesized Mean Difference	0	
Z	-1.364433245	
P(Z<=z) one-tail	0.086215625	
z Critical one-tail	1.644853627	
P(Z<=z) two-tail	0.172431251	
z Critical two-tail	1.959963985	

The p values are exact with respect to the NORMSDIST function.

Note that the “P(t<=T) two tail” statement is in error. It should be “P(T≠t) two-tail”.)

Data Analysis: z-Test: Two Sample For Means: Returns correct values.

TESTS OF SIGNIFICANCE ON CORRELATIONS

EXCEL FUNCTIONS

Excel (2000, 2003 and 2007) does not provide any direct functions to determine whether a correlation value obtained from data is significantly different from a hypothesized value. Excel has two functions (FISHER and FISHERINV) that they claim can be used to make such tests of significance. However in this case, Excel's Help (2000, 2003 and 2007) is misleading. Essentially, they never bothered to actually read Fisher {R. A. Fisher in his book "Statistical Methods, Experimental Design, and Scientific Inference" (The 1990 Oxford 14th Edition), discusses the correlation coefficient in chapter 6.}

EXCEL FUNCTION FISHER:

EXCEL HELP DESCRIPTION OF FUNCTION:

Returns the Fisher transformation at x. This transformation produces a function that is normally distributed rather than skewed. Use this function to perform hypothesis testing on the correlation coefficient.

Syntax

FISHER(x)

X is a numeric value for which you want the transformation.

Remarks

- If x is nonnumeric, FISHER returns the #VALUE! error value.
- If $x \leq -1$ or if $x \geq 1$, FISHER returns the #NUM! error value.
- The equation for the Fisher transformation is:

$$z' = \frac{1}{2} \ln \left(\frac{1+x}{1-x} \right)$$

MISINTERPRETATIONS AND MISUSE USE OF FISHER:

On page 200 Fisher gives the formula for a "z" value, The Excel function FISHER duplicates Fisher's equation. The function converts a correlation coefficient calculated from data (the x value) to a transformed variable "z" which Fisher says is approximately normally distributed with a small bias³.

However the z value here is not the same as the z value entered into NORMSDIST for a p value. Help as currently worded implies that the Fisher z value is the same as the NORMSDIST z value. The confusion is that Fisher used z for two different measures. Excel's Help does not state this. A user will get a totally wrong p value if he uses the FISHER z value as an input into NORMSDIST.

³ Fisher's z value here has a bias of $\{\rho/(2*(n-1))\}$ where rho is the true population correlation coefficient (p 207). The bias is small and can be neglected most of the time.

To correctly use Fisher's method, you have to divide Fisher's z value (after subtracting the bias) by $1/\sqrt{(n-3)}$ (the standard error of z) where n is the number of pairs used to calculate the correlation coefficient (see pages 204 and 205 of Fisher). The resulting value can now be treated as a standard normal distribution z value to obtain a p value (from NORMSDIST). The p value here is the probability that r is different from zero.

Note that this is not the recommended test of whether a sample correlation coefficient comes from a population having a zero correlation (a test for x being significantly different from zero). The recommended test is a t test.

EXCEL FUNCTION FISHERINV

EXCEL HELP DESCRIPTION OF FUNCTION:

Returns the inverse of the Fisher transformation. Use this transformation when analyzing correlations between ranges or arrays of data. If $y = \text{FISHER}(x)$, then $\text{FISHERINV}(y) = x$.

Syntax

FISHERINV(y)

Y is the value for which you want to perform the inverse of the transformation.

Remarks

- If y is nonnumeric, FISHERINV returns the #VALUE! error value.
- The equation for the inverse of the Fisher transformation is:

$$x = \frac{e^{2y} - 1}{e^{2y} + 1}$$

MISINTERPRETATIONS AND MISUSE USE OF FISHERINV:

The same problem that occurs in the use of FISHER, also occurs here. The y value that should be imputed is Fisher's z value, not the normal distribution z value. Also the imputed y value should have the correction for bias that Fisher recommends.

APPROPRIATE TESTS ON CORRELATION COEFFICIENT VALUES

The scope of testing according to Fisher involves a decision of just what is being tested:

1. TESTING IF THE CORRELATION COEFFICIENT FROM A SAMPLE IS FROM A POPULATION THAT HAS A ZERO CORRELATION BETWEEN PAIRS.

The appropriate test is the t-test, not Fisher's z test. Fisher shows how a difference between a sample r value and a proposed population **ZERO** value can be tested for by using the standard t test

$$t = \sqrt{(n-2)} * r / \sqrt{(1 - r * r)} = r * \sqrt{(n-2)} / (1 - r^2)$$

$$df = n-2,$$

$$p = \text{TDIST}(t, df, \text{"1 or 2 tails"})$$

where n is the number of pairs observed.

If the sign of r is unknown (either + or -), then a two-tailed test is appropriate. If the sign is known, then a one-tailed test is appropriate.

Excel's Help should show how to do this. Excel should provide a new function that does this.

2. TESTING IF THE CORRELATION COEFFICIENT FROM A SAMPLE IS FROM A POPULATION HAVING AN EXPECTED NON-ZERO CORRELATION COEFFICIENT BETWEEN PAIRS

In terms of existing Excel functions:

$$z1 = \text{FISHER}(\text{"r of sample"})$$

$$z2 = \text{FISHER}(\text{"expected r value"})$$

$$zd = \text{ABS}(z1 - z2) * \text{SQRT}(\text{"n of sample"} - 3)$$

$$p = \text{NORMSDIST}(zd)$$

For a test of non-equality, use the 2-tailed test.. For a test of either < or a test of > use the one-tailed probability. N is the number of pairs.

Excel's Help should show how to do this. Excel should provide a new function that does this.

3. TESTING IF THE CORRELATION COEFFICIENTS FROM TWO SEPARATE SAMPLES ARE FROM THE SAME POPULATION HAVING AN UNKNOWN CORRELATION COEFFICIENT BETWEEN PAIRS

In terms of existing Excel functions:

$$z1 = \text{FISHER}(\text{"r of sample 1"})$$

$$z2 = \text{FISHER}(\text{"r of sample 2"})$$

$$se = (1/(\text{"n of sample 1"} - 3)) + (1/(\text{"n of sample 2"} - 3))$$

$$zd = \text{ABS}(z1 - z2) / \text{SQRT}(se)$$

$$p = \text{NORMSDIST}(zd)$$

For a test of non-equality, use the 2-tailed test.

Excel's Help should show how to do this. Excel should provide a new function that does this.

4. COMBINING THE SAMPLE CORRELATION COEFFICIENTS FROM TWO SEPARATE INDEPENDENT SAMPLES HYPOTHESIZED AS BEING FROM THE SAME

POPULATION, IN ORDER TO ESTIMATE THE POPULATION CORRELATION COEFFICIENT BETWEEN PAIRS

In terms of existing Excel functions:

$$z1 = \text{FISHER}(\text{"r of sample 1"})$$

$$z2 = \text{FISHER}(\text{"r of sample 2"})$$

$$nt = \text{"n of sample 1"} + \text{"n of sample 2"}$$

$$zave = ((\text{"n of sample 1"} - 3) * z1 + (\text{"n of sample 2"} - 3) * z2) / nt$$

$$r = \text{FISHERINV}(zave)$$

Excel's Help should show how to do this. Excel should provide a new function that does this.