

Problems with Using Excel for Statistical Analyses

**I. Elaine Allen, John D. McKenzie, Jr.,
and William H. Rybolt**

**Math/Science Division
Babson College
Babson Park, MA 02457-0310**

There are numerous reasons to be concerned when considering Microsoft Excel for statistical analyses: for example, the reliability of its statistical procedures. This paper identifies 21 major problems in the latest release of this popular spreadsheet. Unfortunately, these problems will probably continue to exist in future releases based upon Microsoft's past failure to correct such defects.

**Dinner Meeting of the Rhode Island Chapter
of the American Statistical Association
Bryant College
1 March 2001**

Twenty-One Problems with Excel

- 1. Missing Data**
- 2. Importing Web Numbers**
- 3. Hidden Cells and Cell Formatting**
- 4. Histogram**
- 5. Pie Chart**
- 6. Pyramid Charts**
- 7. Mode**
- 8. Display of Digits**
- 9. Variability of Data**
- 10. Quartiles**
- 11. Permutations and Combinations**
- 12. Probability Distributions**
- 13. Bivariate Standardized Normal Distribution**
- 14. Confidence “Interval”**
- 15. Two Independent Sample t Test**
- 16. Paired t Test**
- 17. Coefficient of Determination**
- 18. Multiple Regression**
- 19. Unreliable Algorithms**
- 20. Rank Correlation Coefficient**
- 21. Chi-square**

Missing Data

In statistical calculations in contrast to numerical calculations, a blank cell sometimes behaves as a zero and sometimes as a missing value. The cells (1, blank cell, 2) yield a sum of 3, an average of 1.5, and a standard deviation of 0.7070, while the cells (1, 0, 2) yield a sum of 3, an average of 1, and a standard deviation of 1.

Importing Web Numbers

Import 10, 12, 14, and 15 from a web page.

Enter 5, 6, 7, and 9 by typing.

Calculate the mean and standard deviation of the eight numbers.

Why do students get a variety of values?

format

X1	X2
10	10
12	12
14	14
15	15
5	5
6	6
7	7
9	9
Average 9.75	6.75
Stdev 3.69362385	1.707825128

format spreadsheet as text strings

Hidden Cells and Cell Formatting

When you hide cells in Excel, they are still included in the calculations even though they are not visible. If you copy a range with hidden cells and paste the values, you end up with the hidden cells appearing in the output range.

When you copy and paste, there is no simple way to exclude the hidden cells.

Histogram

Data 100 standardized values

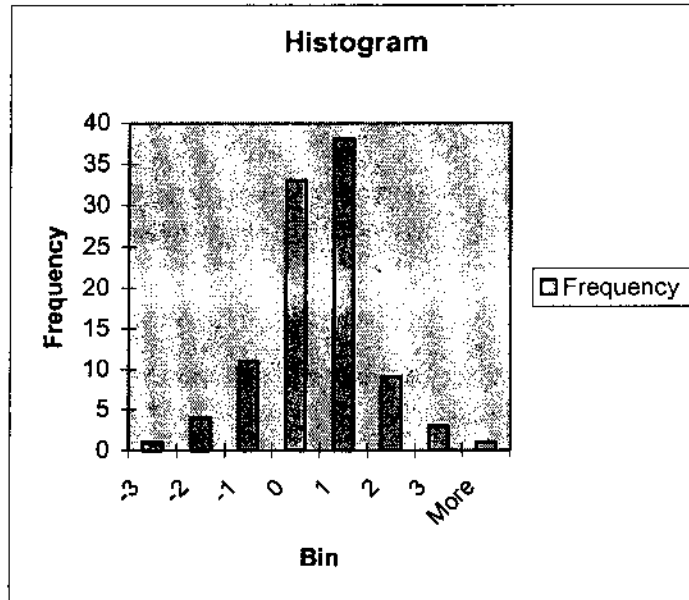
-3.02301	0.16007	-0.86578	0.87326	0.21472	-0.05047	-0.38453
1.25888	0.92624	0.66384	-0.93806	1.07558	0.55485	0.03393
-0.51294	-2.17049	-1.33216	-0.34658	-1.04803	-0.97050	2.27426
-0.13655	-1.17958	-2.59941	-2.36927	-0.31105	0.07945	0.17939
0.25792	0.27194	-0.96915	0.42079	-0.12367	-0.37960	-1.58012
0.27329	0.78345	0.85097	0.04986	-0.51880	1.15509	0.60279
1.70498	1.44462	0.09881	-1.06987	-0.09519	-0.72141	1.09044
-0.80781	0.77103	1.00959	2.72608	3.42727	0.28323	-0.27297
-0.62431	-0.53224	0.99505	-1.98193	-0.31636	-1.32477	0.45990
-1.58736	-2.37436	1.39698	-0.59510	-0.60408	0.22217	0.49921
0.98469	0.60224	-0.03776	-0.82263	0.65491	-0.07914	-0.10557
1.63673	0.65709	-1.23573	0.25088	-0.29176	0.62803	0.27663
0.18250	-0.35488	-1.02979	-0.75507	-0.18873	0.85196	0.04289
2.32598	0.67937	-0.39733	-0.95820	0.40513	-0.02031	-1.52572
0.52371	0.01205					

Excel frequency table:

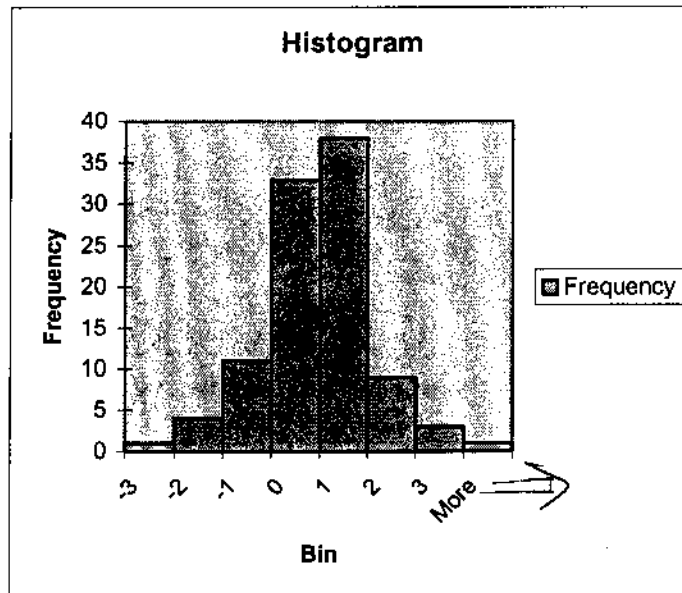
Bin	Frequency
-3	1
-2	4
-1	11
0	33
1	38
2	9
3	3
More	1

Histogram

Excel bar chart:



Excel histogram:



The left-hand tick marks of each interval are really the right-hand tick marks.

Pie Chart

A survey yields of 20 responses with both numerical codes and textual phrases.

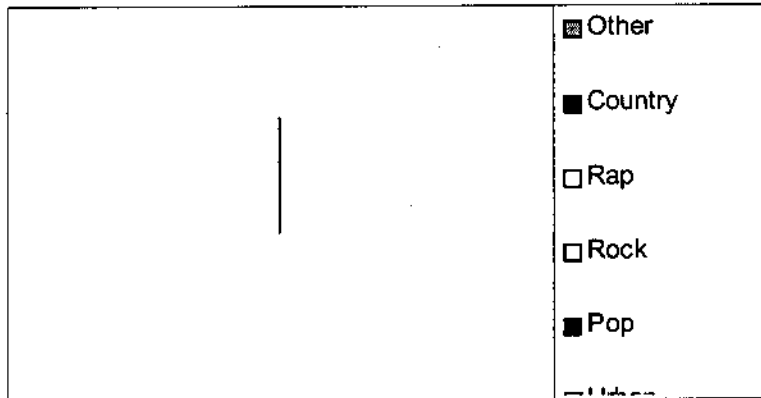
Music Preference	
Code	Category
9	Other
2	Country
5	Rap
1	Rock
4	Pop
3	Urban- Contemporary
1	Rock
3	Urban- Contemporary
8	Jazz
1	Rock
9	Other
9	Other
3	Urban- Contemporary
1	Rock
5	Rap
3	Urban- Contemporary
3	Urban- Contemporary
5	Rap
1	Rock
9	Other

raw data

*create
summary
data first
before using
pie*

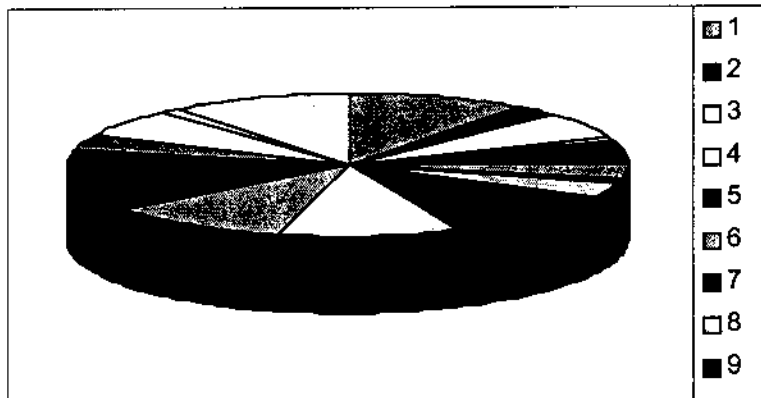
Pie Chart

Pie Chart for the Textual Data



There are no warning messages.

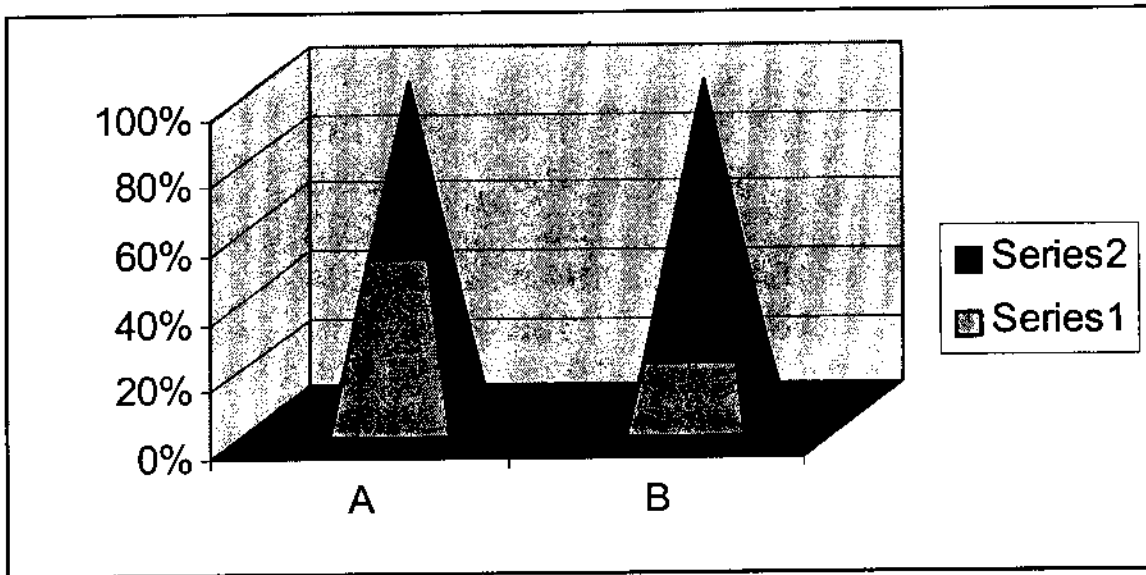
Pie Chart for Numerical Codes



The raw data was not converted into summary data. This is undocumented.

Pyramid Charts

Excel 100% Stacked Column
with a Pyramid Shape Graph



height or volume?

↑
used

Mode

Data ordered lowest to highest, mode = 183.

Data ordered from largest to smallest, mode = 186.

Data order in a random way, mode = 184.5

109	184.5	186
109.5	184.5	188.5
174	185	194.5
178.5	186	195
183	186	202.5
183	188.5	203
184.5	194.5	204
184.5	195	214
185	202.5	109
186	203	109.5
186	204	174
188.5	214	178.5
194.5	109	183
195	109.5	183
202.5	174	184.5
203	178.5	184.5
204	183	185
214	183	186
Mode	Mode	Mode
183	184.5	186

multiple modes

The first number is considered the modal value even if there are multiple modes.

Display of Digits

For example, consider the output from a regression involving the explanatory integers (1, 2, 3, 4, 5, 6, and 7) and corresponding response integers (18, 17, 22, 22, 24, 27, and 27), then Excel reports the Standard Error, t statistic (t Stat), P-value, and other values with an accuracy of nine digits.

Partial Excel Data Analysis ToolPak Regression Output
Presenting a False Sense of Accuracy

Standard Error	t Stat	P-value
1.066656037	14.46442985	2.84977E-05
0.238511541	7.337171166	0.000737717

Variability of Data

When large constants are added to the nine integers {1, 2, 3, 4, 5, 6, 7, 8, 9} the variance and standard deviation change for large constants. Adding 90000000 yields

1	90000001
2	90000002
3	90000003
4	90000004
5	90000005
6	90000006
7	90000007
8	90000008
9	90000009

The mean, standard deviation, and variance:

Mean	5	Mean	90000005
Standard Deviation	2.73861	Standard Deviation	2.82842712
Sample Variance	7.5	Sample Variance	8

should be the same

→ 2788

→ 5

There are no warning messages.

rounding troubles

Quartiles

Find the upper and lower quartiles for a simple set of 11 numbers. Excel gives $Q1 = 35$ and $Q3 = 85$

Position	Value
1	10
2	20
3	30
4	40
5	50
6	60
7	70
8	80
9	90
10	100
11	110

$Q1$ is at position $= 1*(n+1)/4 = 3$. Thus $Q1 = 30$.

$Q3$ is at position $= 3*(n+1)/4 = 9$. Thus $Q3 = 90$.

Why the difference?

Permutations and Combinations

Find the number of ways one can choose 20 objects from 120 when order is important and when order is not important.

The correct answer for the number of permutations is found in statistical functions. The number of combinations is only found in Math & Trig functions.

The output from the permutation function:

PERMUT

Number 120 = 120

Number_chosen 20 = 20

= 7.16787E+40

Returns the number of permutations for a given number of objects that can be selected from the total objects.

Number_chosen is the number of objects in each permutation.

Formula result = 7.16787E+40

OK Cancel

The output from the combination function:

COMBIN

Number 120 = 120

Number_chosen 20 = 20

= 2.94622E+22

Returns the number of combinations for a given number of items. See Help for the equation used.

Number_chosen is the number of items in each combination.

Formula result = 2.94622E+22

OK Cancel

in math's function

Probability Distributions

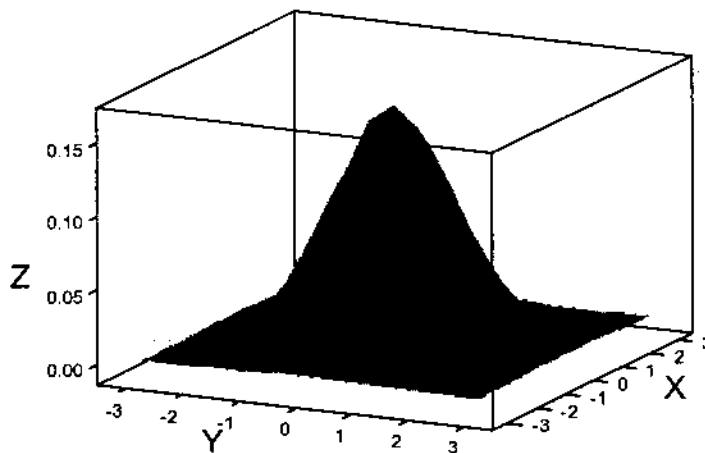
“TDIST is calculated as $TDIST = p(x < X)$, where X is a random variable that follows the t-distribution.”

“NORMSDIST ... Returns the standard normal cumulative distribution function. The distribution has a mean of 0 (zero) and a standard deviation of one. Use this function in place of a table of standard normal curve areas.”

non-consistent

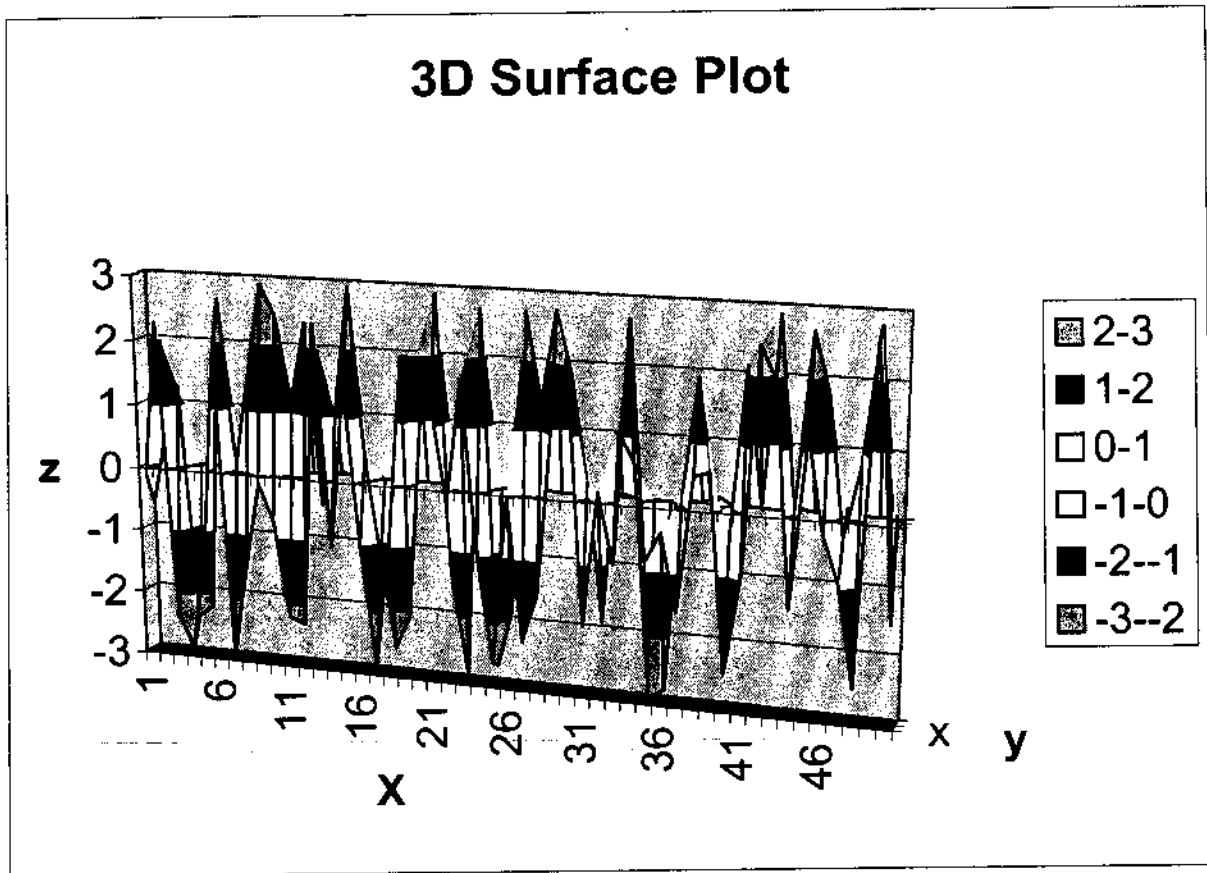
Bivariate Standardized Normal Distribution

Minitab 3D Representation of Bivariate Standardized Normal Distribution



Bivariate Standardized Normal Distribution

Excel 3D Representation of Bivariate Standardized Normal Distribution



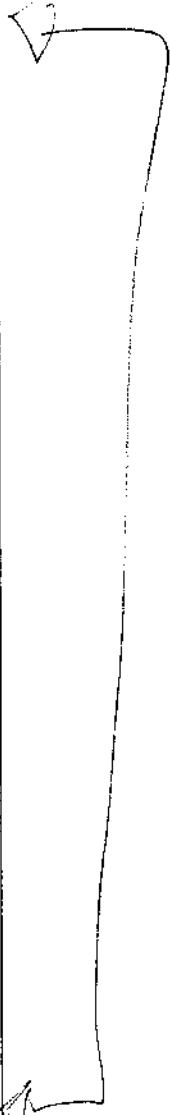
Confidence "Interval"

Excel Output Producing Different Results
for Same Statistical Function

Annotated CONFIDENCE Function Output	
Confidence Interval Function	6.147563548
Data Analysis ToolPak Output	
Mean	33
Standard Error	3.136573806
Median	31
Mode	#N/A
Standard Deviation	12.14789811
Sample Variance	147.5714286
Kurtosis	-1.013153254
Skewness	0.440527085
Range	39
Minimum	17
Maximum	56
Sum	495
Count	15
Confidence Level(95.0%)	6.727287728

not documented $\frac{1}{2}$ width

*one
one
Z₁*



Two Independent Sample t Test

From the 'Help' menu choose 'Contents and Index' and go to the book 'Analyzing Statistical Data.' Within that book, click on 'Perform a t-Test analysis' and accept the first of three choices – 'Learn about the t-Test: Two-Sample Assuming Equal Variances Analyses.'

The first [three] sentences read as follows: 'This analysis tool performs a two-sample Student's t-test. This t-test form assumes that the means of both data sets are equal; it is referred to as a homoscedastic t-test. You can use t-tests to determine whether two sample means are equal.'

documentative
Comparing sample & population

Paired t-test

A Data set has two missing values.

There are three ways to do the paired t-test:

- (1) include the blank, missing cells in his analysis;
- (2) eliminate blank cells by shifting the data up in each column and;
- (3) eliminate the students with incomplete data.

Results when blank cells are included

t-Test: Paired Two Sample for Means		
	<i>Exam1</i>	<i>Exam2</i>
Mean	85.25806452	88.25806452
Variance	27.06451613	57.5311828
Observations	31	31
Pearson Correlation	0.405885548	
Hypothesized Mean Difference	0	
df	30	
t Stat	-0.537530947	
P(T<=t) one-tail	0.297433071	
t Critical one-tail	1.697260359	
P(T<=t) two-tail	0.594866141	
t Critical two-tail	2.042270353	

Paired t-test

Results when the cells are shifted up:

t-Test: Paired Two Sample for Means		
	<i>Exam1</i>	<i>Exam2</i>
Mean	85.25806452	88.25806452
Variance	27.06451613	57.5311828
Observations	31	31
Pearson Correlation	0.105538949	
Hypothesized Mean Difference	0	
df	30	
t Stat	-1.912646539	
P(T<=t) one-tail	0.032691167	
t Critical one-tail	1.697260359	
P(T<=t) two-tail	0.065382334	
t Critical two-tail	2.042270353	

Results when using only complete data

t-Test: Paired Two Sample for Means		
	<i>Exam1</i>	<i>Exam2</i>
Mean	85.48275862	88.5862069
Variance	27.04433498	56.89408867
Observations	29	29
Pearson Correlation	0.405885548	
Hypothesized Mean Difference	0	
Df	28	
t Stat	-2.315480216	
P(T<=t) one-tail	0.014065863	
t Critical one-tail	1.701130259	
P(T<=t) two-tail	0.028131727	
t Critical two-tail	2.048409442	

This last analysis gives the correct results.

Coefficient of Determination

Compute the value of the coefficient of determination (R-squared) for the following set of response and explanatory variables, that is due to Eakin (1996):

Response and Explanatory Values

1.1	10000000.1
1.9	10000000.2
3.1	10000000.3
3.9	10000000.4
4.9	10000000.5
6.1	10000000.6

Excel's RSQ function yields 2.092810987 > 1

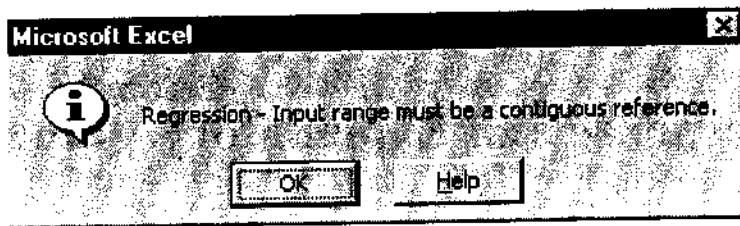
Excel's Regression Analysis Tool yields -0.816484141. < 0

Values must be between 0 and 1.

The correct answer is 0.997.

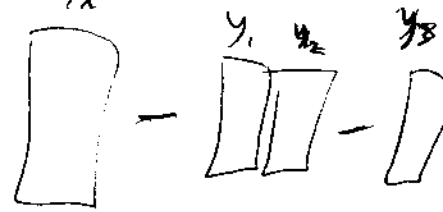
Multiple Regression

Excel Error Message for Noncontiguous Data



You must make columns

to be contiguous



Unreliable Algorithms

Partial Excel Data Analysis ToolPak
Output for Collinear Data

Regression Statistics	
Multiple R	65535
R Square	-0.460636874
Adjusted R Square	-3.65197E-09
Standard Error	0.000142846
Observations	9

2¹⁶

65535

0/0

←

Rank Correlation Coefficient

Compute the value of the rank correlation coefficient for

Test 1	Test 2
93	18.0
83	15.0
90	2.0
60	5.0
25	7.5
50	20.0
94	23.0
99	16.0
62	9.0
97	4.0
43	11.5
95	18.0
84	21.0
79	14.0
62	24.0
100	7.5
83	11.5
85	11.5
52	11.5
100	6.0
100	22.0
25	1.0
84	3.0
41	18.0

Excel yields 0.075593; the correct answer is 0.101.

Excel handles ranks as standings are reported in the world of sport. For example, if two teams are tied for second each receives a rank of 2. *(rather than 2.5)*
each

Chi-square

	A	B	C	D	E	F	G	H
2		Actual Values						
3		Summer	Winter					
4	Yes	163	154	317				
5	No	64	108	172				
6		227	262	489				
7								
8		Expected Values						
9		Summer	Winter					
10	Yes	147.1554192	169.8445808	317				
11	No	79.84458078	92.15541922	172				
12		227	262	489				
13								
14	p-value	0.002623218						
15								
16								
17								

10:C11

Output

Actual_range: B4:C5 = 163,154,317

Expected_range: B10:C11 = 147.1554192,169.8445808

= 0.002623218

Returns the test for independence: the value from the chi-squared distribution for the statistic and the appropriate degrees of freedom.

Expected_range is the range of data that contains the cells of the product of row totals and column totals in the grand total.

Formula result: =0.002623218

OK Cancel

↑
we have tail