Paper SP06

On the use of spreadsheets in statistical analysis

Martin Gregory, Merck Serono, Darmstadt, Germany

1 Abstract

While most of us use spreadsheets in our everyday work, usually for keeping track of lists, very few statisticians or statistical programmers in clinical research use them to do statistical computations. Outside of clinical research, however, many people do use the statistical functions in spreadsheets for statistical analysis. There is a large body of literature, both in journals and on the internet studying the issues which arise when using spreadsheets for this purpose. In addition to factors common to standard statistical software such as numerical accuracy and correctness of implementation, the execution model employed by spreadsheets presents special challenges and makes it legitimate to ask whether spreadsheets are a suitable tool for such analyses. This paper presents a survey of the findings and summarizes the recommendations on the use of spreadsheets for statistical analysis.

2 Introduction

Ask a collection of statistical programmers or statisticians from the pharmaceutical industry to name the most commonly used software for statistical analysis and the answer will, most likely, be SAS. The most commonly *installed* software for statistical analysis is, however, almost certainly Microsoft's Excel software. According to Brian Ripley [1], speaking at the Royal Statistical Society conference of 2002:

> Let's not kid ourselves: the most widely used piece of software for statistics is Excel.

There are popular books on using Excel for statistical analysis [2] or for data analysis, including statistics, in a scientific environment [3] which support Ripley's assertion. Web sites devoted to the topic, for example Arshan's [4] may also be found.

Microsoft's *Data Analysis Toolpak for Excel* is provided by default with Excel but needs to be enabled before it can be used. This package provides descriptive statistics, various statistical tests (e.g. t-test, paired t-test, χ^2), calculation of correlation matrices, ANOVA and various regression analyses. Furthermore, Excel include many functions for statistical distributions and a random number generator. The functions are enabled by default.Microsoft's Excel is not the only spreadsheet with these capabilities. The open source software packages Gnumeric and Open Office Calc both provide similar features although Open Office Calc is moving towards providing an interface to R as the primary approach to statistical analysis.

Given the existence of these software tools which have either zero cost or already paid for licences, it is legitimate to ask whether these tools can augment or indeed replace the current collection of statistical analysis tools. Are they easy to use? Do they easily lend themselves to generalization and automation? Do they produce correct results? Finally, and especially of interest to programmers and statisticians in the pharmaceutical industry, do they facilitate reproducibility and tracking of results, features of utmost importance for regulatory compliance. In attempting to answer these questions, we restrict ourselves to the most commonly used version Excel 2003, with some references to literature on earlier [12, 13, 16, 17] and later [19, 15, 11] versions. Comparison with Open Office Calc was done using version 2.4.0.

3 The spreadsheet execution model and its consequences

Although we probably do not think of it so, a spreadsheet is a rather specialized type of programming language. Like other programming languages it takes inputs, performs calculations on them and produces outputs. For certain types of tasks it is eminently suitable. To determine its suitability for statistical analysis,we consider a number of ways in which it differs from programming languages or packages normally used for this purpose:

1. there is no separation of program code (formulas) and data;
2. the basic unit of a spreadsheet, the cell, may contain either data, a reference to another cell or program code;

3. although this need not necessarily be so, all the spreadsheets we consider exhibit dynamic data typing. Unlike type casting in other programming languages, however, the typing is based not on context but on the value entered;

4. unless otherwise specified, modifying a cell results in the recalculation of all formulas present, i.e. the program is executed again;

5. recalculation of a cell containing a formula may change its displayed value;

6. unless explicitly protected, any cell may be modified by the user;

7. results are generally written to a cell or range of cells, possibly in another sheet or workbook.

The lack of separation between program code and data makes re-use of the code a manual exercise as a different set of data needs to be loaded into the spreadsheet. This clearly increases the risk of human error. Additionally, it is generally the case that different data set have different dimensions. By default, formulas which work on ranges do not automatically modify ranges according to the data present. While this is possible, the setting is global, so analyses which do require a fixed dimension are liable to error based on providing incorrectly dimensioned data.

The lack of obvious distinction as to the content of cells may cause unforeseen results. For example, it is possible to sort on a column or range containing formulas and this may produce a different result from sorting a column or range containing, as constants, the results of the formulas. For example, create a 1×8 vector, each cell containing the formula *RAND()*. In our test, this produced the following values:

$$0.64 \quad 0.76 \quad 0.91 \quad 0.57 \quad 0.01 \quad 0.89 \quad 0.70 \quad 0.94$$

Sorting this range with the auto calculate option turned on, the default, we obtained:

$$0.85 \quad 0.03 \quad 1.00 \quad 0.69 \quad 0.25 \quad 0.79 \quad 0.45 \quad 0.05$$

i.e., in the process of sorting, the formula was re-executed. Turning off the auto calculate option leaves the order unchanged. In both cases, the spreadsheet is sorting the formula itself, not the results[1].

Not only is this recalculation done during a sort: if the auto calculate is turned on, changing any cell, or even simply opening the spreadsheet will trigger the recalculation.

4 Input handling

Unlike the standard programming model, the input data are not stored in a separate file but rather in a range of cells within the spreadsheet. Formulas or the *Data Analysis Toolpak* analyses reference this range of cells. When an analysis needs to be run on a different set of data, the input range may grow or shrink. By default, ranges do not extend to cover inserted cells: they need to be manually changed. Both Excel and Open Office Calc have options to automatically re-size ranges, but these are global options so either all ranges re-size or none do. There is a significant possibility that ranges may be set incorrectly when switching data so care must be taken to ensure that this does not happen. The act of replacing data is itself a possible source of error. In comparison to the normal programming language paradigm where input data are stored in separate files, the spreadsheet model is time-consuming and error-prone.

5 Output handling

A further consequence of the execution model pertains to the placing of output which must be almost entirely handled by the user. This is another aspect of the execution model which makes generalization difficult unless one resorts to a scripting language such as VBA for Microsoft Excel or Python for Open Office Calc.

Output produced by functions are written to the cell containing the function, so this cell needs to be placed in the output area. If analysis by factors or by groups is required, this requires manual intervention. For the *Data Analysis Toolpak* output ranges must be manually specified by the user but may be a range in an existing sheet or a new sheet. Analysis by groups can be done with the help of Pivot Tables. With some exceptions in ATP, labelling of output must be done manually.

In summary, much of the layout which is done automatically in a standard statistical package such as SAS or R must be done manually for each analysis. Additionally, there is no structured output object concept as in R or SAS, so there is no easy way of passing results of one analysis to a later one.

[1] Repeating this experiment will produce different results as the function is seeded by the current time

6 Execution tracking

Although not a consequence of the spreadsheet execution model, we have not encountered any feature for reporting on what code was executed in any of the spreadsheets examined, i.e. there is no equivalent of the SAS log or equivalent features in other standard statistical packages. The SAS log provides valuable information when debugging programs but, more importantly for statistical programming in the pharmaceutical industry, it provides a record of what actions were carried out. It generally provides useful information on the number of records read or written or used in a calculation. The lack of this feature is one of the most serious objections to the use of spreadsheets for statistical analysis in a regulated setting.

7 Two examples of the Data Analysis Toolpak

We now present two examples of the use of the *Data Analysis Toolpak* in order to provide an impression of the steps required to carry out analyses. The examples are taken from a very informative paper by Goldwater [5] which paper contains further examples of other analyses. The analyses use the following fictitious data shown in Table 1.

Table 1: Sample data

Treatment	Outcome	X	Y
A	1	10.2	9.9
A	1	9.7	
B	1	10.4	10.2
A	2	9.8	9.7
B	1	10.3	10.1
A	2	9.6	9.4
B	1	10.6	10.3
A	2	9.9	9.5
B	2	10.1	10
B	2		10.2

The first analysis is to produce means and standard deviations of X and Y for the entire data and for each treatment group. The second is to carry out a paired t-test to determine whether the means of X and Y show a statistically significant difference.

7.1 Descriptive statistics

To calculate the mean and standard deviations of X and Y, we can select the *Descriptive Statistics* option from the *Tools, Data Analysis* menu. In the dialog window we may choose the input range, in this case the columns containing X and Y. Empty cells are ignored. By including the label rows, the output will be labelled and will contain the results shown in Table 2.

Note that in order to perform the analysis on multiple columns, these must be adjacent, i.e. the *Data Analysis Toolpak* does not accept non-contiguous selections. This greatly increased the amount of manual work required when carrying out analyses on multiple variables.

In order to perform the analysis by treatment group it is necessary to either re-arrange the data so that all observations for a group are contiguous and then use the *Data Analysis Toolpak.* Alternatively, if re-arrangement is not desirable and if the Pivot Table feature provides the required statistics, by using a Pivot Table. This involves a number of drag and drop actions equal to the product of the variables being analysed and the statistics required and rapidly becomes very tedious with increasing numbers of variables or groupings.

7.2 Paired t-test

To calculate the paired t-test for difference in means of X and Y by subject, we can select the *t-Test: Paired Two Sample for Means* option from the *Tools, Data Analysis* menu. In the dialog window we may choose the two variables independently, i.e., these may be non-adjacent. We must also enter the hypothesised difference for which there is no default. By including the label rows, the output will be labelled and will contain the results shown in Table 3.

Table 2: Descriptive statistics

X		Y	
Mean	10.06666667	Mean	9.922222222
Standard Error	0.113038833	Standard Error	0.107726219
Median	10.1	Median	10
Mode	#N/A	Mode	10.2
Standard Deviation	0.339116499	Standard Deviation	0.323178657
Sample Variance	0.115	Sample Variance	0.104444444
Kurtosis	-1.171482582	Kurtosis	-1.087022247
Skewness	0.120884088	Skewness	-0.598984565
Range	1	Range	0.9
Minimum	9.6	Minimum	9.4
Maximum	10.6	Maximum	10.3
Sum	90.6	Sum	89.3
Count	9	Count	9

Table 3: Paired t-test results

	X	Y
Mean	10.06666667	9.922222222
Variance	0.115	0.104444444
Observations	9	9
Pearson Correlation	0.950659175	
Hypothesized Mean Difference	0	
df	8	
t Stat	0.087010555	
P(T¡=t) one-tail	0.466400792	
t Critical one-tail	1.859548033	
P(T¡=t) two-tail	0.932801584	
t Critical two-tail	2.306004133	

From the data, where every value of X is greater than the corresponding Y value, we would expect a lower p-value. If we compare the results to those from R we find a significant discrepancy in both the t statistic and the 2-tailed p-value:

```
Paired t-test

  data:  X and Y
  t = 6.1482, df = 7, p-value = 0.0004684
  alternative hypothesis: true difference in means is not equal to 0
  95 percent confidence interval:
   0.1384636 0.3115364
  sample estimates:
  mean of the differences
                    0.225
```

We notice that R gives degrees of freedom as 7 while *Data Analysis Toolpak* gives 8. On further examination we see that Excel did not exclude the observations with a missing value for one of the pair. If we manually exclude these two observations, Excel produces the correct result. So Excel does not handle observations with missing values for one of the pairs correctly.

Curiously enough, there is a function, *TTEST()*, which gives the correct result for this task, but being a function it returns only a single value: the significance of the 2-tail test. No other information is returned.

8 Further considerations

In this section we discuss several factors relevant to our topic which are not related to the spreadsheet execution model but appear to be common features of the spreadsheet software we examine.

8.1 Missing values

Missing values are generally ignored in calculations. For example, applying the *AVERAGE()* function to the vector

$$1 \quad 2 \quad . \quad 3$$

produces the result 2, i.e. using a denominator of 3 for the 3 non-missing items. However, a reference to a cell containing a missing value returns 0, so a range containing references to the above ranges shows

$$1 \quad 2 \quad 0 \quad 3$$

and returns a mean of 1.5, i.e. using a denominator of 4.

8.2 Precision in saving ASCII files

In general spreadsheets operate with double precision floating point arithmetic. Both Excel and Open Office Calc provide global options which cause calculations to be done using the precision of the displayed values. By default these are switched off and, in general, should not be switched on as not only do they use reduced precision in calculations, they actually change the precision of constant values in the spreadsheet, i.e. the input values are changed.

As Burns [6] has pointed out, in the particular case of writing tab or comma separated files, Excel defaults to using precision as displayed, i.e. the full value will not be written to the text file. This means that care must be taken not to format numeric cells when a spreadsheet is being used as a transfer medium. Open Office Calc presents an option to save with precision as displayed, but this is not the default.

8.3 Table size

Both Excel and Open Office Calc allow only 64k (2^{16}) rows and 256 columns. While this may suffice for many purposes, the row limit is especially low. For example, laboratory data for large Phase III trials easily exceed this limit. Excel 2007 allows about one million (2^{20}) rows and 16k (2^{14}) columns.

8.4 Unary minus

The unary minus operator has a higher priority than other operators: calculating -3^2 in Excel or Open Office Calc returns 9. SAS like most programming languages returns 9. Interestingly enough, VBA, which may be used to extend the functionality of Excel, also returns 9 [6]. While there is no a priori reason to choose one rule over the other, this particular design choice in Excel can lead to inadvertent mistakes.

9 Accuracy

We now turn to the question of how spreadsheets perform as far as the accuracy of statistical features is concerned. This is a topic of interest for any statistical software as is shown by the abundance of literature on the subject. Sawitzki proposed one approach [7] and applied it to various statistical software packages [8]. McCullough [9] proposed a methodology covering three different areas: statistical distributions, estimation and random number generators. The first are assessed using a program known to produce correct results such as Knüsel's [10] ELV program. Estimation is assessed using a set of Statistical Reference Datasets produced by the American National Institute Standards and Technology [2] for which correct results for estimation in four areas, univariate summary statistics, one-way ANOVA, linear regression and non-linear regression. For random number generators, various empirical tests of randomness are used. See McCullough [11] for details of why the generator in Excel 2003 and 2007 is a poor implementation.

[2] http://www.itl.nist.gov/div898/strd/

9.1 Statistical Distributions

The calculation of values across the domain of a statistical distribution generally requires an algorithm which recognizes when the calculation has converged and a result has been found or when it will fail to converge and no result can be reported. Each version of Microsoft's Excel containing functions for calculating statistical distributions has been tested on a variety of distributions. Knüsel [12] examined Excel 97 and found it wanting to such an extent that he recommended against using Excel for statistical purposes. All issues found were reported to Microsoft. Following the release of Excel 2000 and Excel XP, Knüsel [13] and McCullough and Wilson [17] found that none of the reported problems had been fixed. Knüsel [14] repeated the analysis for Excel 2003 finding that while there had been attempts to fix the problems, the solutions exhibited other, sometime even worse, problems. For example, the calculation of Poisson and Binomial distributions previously returned correct values in the tails and incorrect values in the central regions while Excel 2003 simply reversed the situation. The assessment of Excel 2007 was carried out by Yalta [15] who found very little change from Excel 2003. Yalta also examined Gnumeric 1.7.11 and Open Office Calc 2.3.0. Gnumeric produces exact results for all tested distributions with the exception of very small values for the inverse t distribution. Calc returns incorrect values on parts of the domains for inverse χ^2 and inverse β. It is also unable to calculate values for inverse t and inverse F for very small values although it does, at least, report this fact in comparison to Excel which simply returns an invalid result. Judged on the accuracy of functions calculating statistical distributions, Gnumeric is the only spreadsheet which may be recommended.

9.2 Estimation

In the area of estimation, a picture similar to that for statistical distributions emerges. According to McCullough and Wilson, the performance of Excel 97 [16] and Excel 2000/XP [17] was not acceptable. Even a seemingly simple calculation such as the sample standard deviation was extremely unstable, producing zero accurate digits on some tests. For the ANOVA tests, the F statistic was calculated to an accuracy of less than two digits for more than half of the tests. Linear regression tests failed to detect collinearity. Repeating the analysis on Excel 2003 they found significant improvements to the extent that the performance on univariate statistics, ANOVA and linear regression was acceptable. For non-linear regression, however, Excel returns results with zero accurate digits in 21 of the 27 tests [18]. McCullough and Heiser [19] report no changes in the behaviour for Excel 2007.

10 Conclusion

We have seen that while it may be easy to enter a small amount of data and to perform one or two statistical analyses, any more complex analysis on the scale of those performed routinely using standard statistical software such as SAS rapidly becomes time-consuming and error-prone. The research which has been done on the accuracy of statistical features in spreadsheets does not inspire confidence in the ability of spreadsheets to produce the right answers. For statisticians and programmers in the pharmaceutical industry, the lack of traceability and reproducibility is a significant concern. We can only recommend, as the majority of sources quoted do, that spreadsheets are not a suitable tool for statistical analysis.

References

[1] Ripley, B.D. (2002) Retrieved from http://www.stats.ox.ac.uk/ ripley/RSS2002.pdf 28 June 2009

[2] Schmuller, J. (2009) *Statistical Analysis with Excel For Dummies, 2nd Edition* John Wiley & Sons

[3] Ravens, T. (2004) *Wissenschaftlich mit Excel arbeiten* Verlag Pearson Studium

[4] Arshan, H. (2007) *Excel For Statistical Data Analysis* Retrieved from http://home.ubalt.edu/ntsbarsh/excel/excel.htm 21 June 2009

[5] Goldwater, E. (2007) *Using Excel for Statistical Data Analysis - Caveats* Retrieved from http://www-unix.oit.umass.edu/ evagold/excel.html 27 June 2009

[6] Burns, P. (2009) *Spreadsheet addiction* Retrieved from http://www.burns-stat.com/pages/Tutor/spreadsheet_addiction.html, 21 June 2009

[7] Sawitzki, G. (1994) *Numerical reliability of data analysis systems.* Computational Statistics & Data Analysis 18 (2) 269-286

[8] Sawitzki, G. (1994) *Report on the reliability of data analysis systems.* Computational Statistics & Data Analysis 18 (2) 289-301

[9] McCullough, B.D. (1998) *Assessing the reliability of statistical software part I*, Amer. Statist. 52 358-366

[10] Knüsel, L. (1989) *Computergestützte Berechnung Statistischer Verteilungen* Oldenburg, München-Wien (An English version of the program can be obtained at http://www.stat.uni-muenchen.de/˜/knuesel/elv)

[11] McCullough, B.D. (2008) *Microsoft Excel's 'Not The Wichmann-Hill' random number generators* Computational Statistics & Data Analysis 52 (10) 4587-4593

[12] Knüsel, L. (1998) *On the accuracy of statistical distributions in Microsoft Excel 97* Computational Statistics & Data Analysis 26 (3) 375-377

[13] Knüsel, L. (2002) *On the reliability of Microsoft Excel XP for statistical purposes* Computational Statistics & Data Analysis 39 (1) 109-111

[14] Knüsel, L. (2005) *On the accuracy of statistical distributions in Microsoft Excel 2003.* Computational Statistics and Data Analysis. 48 (3) 445-449

[15] Yalta, A.T. (2008) *On the accuracy of statistical distributions in Microsoft Excel 2007* Computational Statistics & Data Analysis 52 (10) 4579-4586

[16] McCullough, B.D. & Wilson, B. (1999) *On the accuracy of statistical procedures in Microsoft Excel 97* Computational Statistics & Data Analysis 31 (1) 27-37

[17] McCullough, B.D. & Wilson, B. (2002) *On the accuracy of statistical procedures in Microsoft Excel 2000 and Excel XP* Computational Statistics & Data Analysis 40 (4) 713-721

[18] McCullough, B.D. & Wilson, B. (2005) *On the accuracy of statistical procedures in Microsoft Excel 2003* Computational Statistics & Data Analysis 49 (4) 1244-1252

[19] McCullough, B.D. & Heiser, D.A. (2008) *On the accuracy of statistical procedures in Microsoft Excel 2007* Computational Statistics & Data Analysis 52 (10) 4570-4578

Contact Information

Martin Gregory
Head of Statistical Programming Germany
Merck KGaA
Frankfurterstr 250
64293 Darmstadt
Germany
Martin.Gregory@merck.de