

II. GENERAL PROBLEMS WITH EXCEL.....	2
SCOPE:.....	2
WORKSHEET ERRORS, FAULTS AND INACCURACIES.....	2
THE PROBLEM.....	2
THE UPDATING – NOT UPDATING ISSUE.....	3
DATA AND EQUATION FALSIFICATIONS.....	4
A VIEW OF THE PROBLEM.....	4
EXCEL INTRINSIC FUNCTION AND ROUTINE FAULTS, ERRORS, DEFICIENCIES AND PROBLEMS.....	4
PROBLEMS, FAULTS, DEFECTS AND ERRORS:.....	5
LEVELS AND STRUCTURE.....	5
WHAT IS A PROBLEM, DEFECT, FAULT OR ERROR?.....	6
PROBLEMS:.....	7
DEFECTS AND ERRORS:.....	7
FAULTS:.....	9
REPORTED GENERAL PROBLEMS WITH EXCEL:.....	9
HELP:.....	10
OVERVIEW:.....	10
FIXES AND COMMENTS:.....	11
STATISTICAL ANALYSIS DOCUMENTATION:.....	12
OVERVIEW:.....	12
MISSING OR DELETED DATA:.....	12
OVERVIEW:.....	12
FIXES AND COMMENTS:.....	13
RANGE AND FUNCTION NAMING REFERENCE ERRORS.....	14
LIMITED STATISTICAL FUNCTIONS AND ROUTINES.....	14
DIFFERENT LANGUAGE VERSIONS OF EXCEL.....	14
EQUATION PRECEDENCE ERRORS.....	15
MIXED TYPE B PROBLEMS AND FAULTS.....	16
THE COMPLEXITY OF EXCEL.....	16
SINGLE-CELL OR ARRAY FORMULA ENTRY.....	16
OVERVIEW:.....	16
FIXES AND COMMENTS:.....	17

DATA INPUT PROBLEMS:	17
THE CELL EQUATION JAMMING PROBLEM.....	19

II. GENERAL PROBLEMS WITH EXCEL

SCOPE:

WORKSHEET ERRORS, FAULTS AND INACCURACIES

THE PROBLEM

An Excel worksheet creates an illusion of orderliness, accuracy and integrity. The rows and columns of data, instant calculations, automatic updating, and other features contribute to this impression. However in business these errors can be disastrous financially and legally. (Panko 2005b)

There are different ways Excel worksheets/workbooks are handled. In business they are commonly passed on to others within the organization or outside, for comments, additions, changes, consolidations, or revisions. Errors, faults and inaccuracies, whether they come from human entry errors or errors from inherent Excel functions or macros, can result in bad business decisions, faulty financial statements and give misleading information about internal planning and tracking.

“Broadly speaking, when humans do simple mechanical tasks, such as typing, they make undetected errors in about 0.5% of all actions. When they do more complex logical activities, such as writing programs (creating a formula, building VBA functions, macros and subroutines), the error rate rises to about 5%. These are not hard and fast numbers, because how finely one defines, “reported action” will affect the error rate.” (Panko 2005b)

In general, even the simplest worksheet calculations aren't properly tested or checked in most companies and in academic research. Untested spreadsheets are riddled with errors, often because the wrong formula was entered in the first place. There are almost no worksheet software quality assurance practices in businesses. (Panko 2000a)

Flawed mathematical models, the huge size and complexity of many financial worksheets, lax security, inordinate ease of access and use, and other factors all contribute to these vulnerabilities, whether through fraud or by accident. Excel 2007 with its major new look and feel and major increase in the allowable size limits (equations, rows, columns) and other visual changes make it even easier to make mistakes. Worksheets built using well-engineered code libraries are inevitably tinkered with later by traders, salespeople, analysts, and other users in an uncontrolled fashion. (Croll, 2005)

There is also the illusion that everything being computed is accurate in the arithmetic sense. There is that prolonged view that the numbers in a worksheet are absolutely correct. This overlooks the errors introduced by the user, the errors and mistakes in the data or preset worksheets and those inherent errors from the software. Both Volpi (2006) and Croll (2005) report that the most errors and faults come from the user side, and the business chains involved in processing and passing data and analysis from organization to organization.

Panko (2005a) discusses the problems of spreadsheet errors and describes methods to find errors, reduce them and ways to manage the quality control process. Although focused on financial spreadsheets where legal requirements exist, the methods he talks about are generally applicable.

Palmer (2006) points out that a significant source of worksheet errors when Excel is used in engineering organizations, is the entry of incorrect equations in cells and in groups of cells.

THE UPDATING – NOT UPDATING ISSUE

One of the fundamental concepts and applications of spreadsheets is that when some part of a “sheet” (cell, input box, etc.) is changed, any other part of the worksheet and other worksheets linked to the given cell or input box changes. This change is called automatic updating, and is managed by changing the options. [For Excel-2003 and earlier versions it is (→ Tools → Options → Calculation). For Excel-2007 it is (Office Button → Excel Options → Formulas → Calculation Options → Workbook Calculation)] All cell equations, formulas and functions are recalculated when the update operation occurs. Normally a change in any cell triggers off automatic updating, unless the manual updating is set.

Nash (2006), talks about this problem and that it can lead to all kinds of hidden errors and faults. He says, “Updating is useful, but it is also dangerous, since we can do a lot of damage with clumsy fingers on the keyboard.” Nash would recommend that the automatic update option be turned off, if “clumsy fingers” is a problem.

The automatic updating (or manual updating via F9) does not change all values on the worksheet. Any macro or subroutine (all the → Tools → Data Analysis → Analysis Tools are subroutines) linked to a sheet **WILL NOT CHANGE ON THE SHEET, UNLESS THE MACRO OR SUBROUTINE IS REDONE** by an effort on the part of the user to re-run that macro or subroutine. Pushing F9 does not redo these subroutines. These outputs can be overlaid if no change is made in the output location.

For example, doing ANOVAs or regression on data on a worksheet, using “→ Tools → Data Analysis → Analysis Tools → Anova: Single Factor” or “→ Tools → Data Analysis → Analysis Tools → Regression”. When these are run, the resulting table of values is a fixed set of cell numbers and text. If the data is now changed, these “outputs” will not change. Consequently results do not reflect the current data set, and this is a source of hidden errors.

If instead of using the “Tools” for regression, one uses the LINEST function (as an array function), the resulting block: of output values (Use Help to find out what each cell is, since no titles or names are output) represent the actual input data “block”. If a number in this input block is changed, the corresponding LINEST output block values will change. That is the update operation changes LINEST outputs when the inputs change. The only change that is blocked is when additional data is to be added, or data is to be removed (the size of the block can’t be changed.). However the internal range changes when columns, rows or cells are added or removed, and the LINEST values do not change.

There are also many other situations where the automatic updating (when a cell changes) does not occur. Pushing the F9 key may do it. Refer to the many KBA's that describe situations where worksheets will not update.

DATA AND EQUATION FALSIFICATIONS

Nash (2006) briefly addresses this problem. The falsification may be accidental or deliberate. Panko (2005a and 2005b) talks about this problem and gives some solutions. In all cases an audit trail must be set up to check initial data inputs, and track all changes made to the data and computations. Audit trails require an administrative procedure, and special software to identify changes made to a given worksheet. Panko (2005a and 2006) talks about administrative procedures. Unfortunately, in science, falsification does occur in deliberate data changes, weak analytical tools, and incorrect interpretations of results.

A VIEW OF THE PROBLEM

In a very general sense, the problem is actually five general problems.

- A. The error, fault or problem was entirely the user's, such as wrong data input or wrong selections of functions and macros, wrong equations, wrong cell references, incorrect fixes and patches, faulty work-arounds, etc. This is the human part of THE PROBLEM discussed above. In general only auditing can detect and correct these errors (Panko 2005b and 2006)
- B. The error, fault is with the user, but Microsoft shares the error/fault because of excessive complexity, misleading information, etc.
- C. The error, fault is entirely the fault of Microsoft. Errors in functions and Data Analysis routines, incorrect directions on usage, wrong or incorrect algorithms, bad error handling, weird displays, obsolete approaches, etc. This includes errors in HELP and deficiencies in Graphics.
- D. Excel does not have the required or needed statistical functions or routines or other needed capabilities. This is a "missing-ness" problem.
- E. The error, fault is in an add-in supplied by a commercial or academic source that has no documented set of tests to verify the accuracy of the add-in. Just about all of them have no trace of testing (data and results) to verify the developer's claims. Microsoft can't be faulted for failures in the add-in developer's software.

EXCEL INTRINSIC FUNCTION AND ROUTINE FAULTS, ERRORS, DEFICIENCIES AND PROBLEMS

The primary purpose of this paper is to focus on reported faults, errors, deficiencies and problems with Excel when used for statistical applications. These are all type C problems. The human part of the problem is not discussed in this paper, since this is well covered elsewhere.

Microsoft has built a large data base of information on their products. This is the "Knowledge Based Articles" or KBA base. It includes all kinds of problems, fixes, workarounds and general information. Note C is a listing of **SOME** of the KBAs regarding Excel's statistical function and routine outputs and problems. This makes it easier to identify faults, errors, deficiencies and fixes.

The Excel statistical functions and routines are used in business applications. A user is unaware when these functions and routines are used that the outputs may have considerable error. The intent in this paper is to identify where and when significant errors occur.

Another purpose is to describe work-arounds and ways to avoid or minimize incorrect Excel outputs.

The third purpose is to see how Excel fits into contemporary introductory statistics courses, where commercial computer programs and software are emphasized. The view is how well does Excel cover the course/text and what are the type D problems.

Some of the problems that can be considered to be type B problems are also covered in this section

In this paper, the focus is on the statistical and related functions and routines (including related graphs) in Excel that would be used to solve statistical problems. The problems with the “outside structure”, that is common to the general use of Excel are not discussed unless it directly affects the use of Excel as a “statistical analysis tool”.

Note E, “Guide To Excel Functions and Tools” is a general listing of those functions and routines that would be used in statistical analysis of data, together with some information of what they do and what are the required inputs. The list is organized into general categories that fit the major topics in most introductory statistics courses.

Add-ins, macros and VBA (Visual Basic for Applications) functions and subroutines incorporated in textbook-supplied CD’s or Internet downloadable *.xla files, were evaluated to a very limited extent (Section 19). They are poorly defined black boxes, with no open literature test results on their accuracy. Excel cannot be faulted if these are faulty. Review and extensive testing of these add-ins is beyond the scope of this article.

PROBLEMS, FAULTS, DEFECTS AND ERRORS: LEVELS AND STRUCTURE

The human user, the computer and the computer software are large complicated structures. From an information-processing standpoint, the human user is at the highest level of this structure, and the computer processor (CPU) is at the lowest level. Between these levels, the structure involves outside people, intelligence and a social/business/academic structure to create software that connects the highest level to the lowest levels. This structure in a very broad sense can be defined as seven levels. These levels would be as follows:

1. How the observed data connects to solving the substantive problem (Mallows 1998). This is basically a plan of the path or steps to solution of the substantive problem.
2. Choice of approach. (Overall View)
3. Choice of method
4. Mathematical and logical expressions.
5. Implementing algorithms (i.e. equations and arithmetic operations)
6. Software language statements. Includes cell equations and links to other cells.
7. CPU instructions and computer processes on these instructions.

At levels 1, 2 and 3, problems, faults, defects and errors can occur, and the human user is the cause. This is a very difficult area to deal with, since it involves opinions by different statisticians, variations in methods given in textbooks (that also misname tests), real life textbook errors, and a variety of possible solution approaches, depending on what article, textbook or authority is used as a basis. This is a real disaster area when it involves tests of a hypothesis.

The next levels are in the computer software, provided from a company (commercial) or from academic institutions. In this group, the design and building of the software involves choices of methods (3) to be included, the specific mathematical expressions (in the literature) to be used (4), a reduction to computer implementations (5), making language (C++, VBA, Fortran, R, IML, etc.) statements (6) and compiling to the CPU instruction sets (7).

Between the upper and lower level, problems, faults, defects and errors, shift from the user to the software. This is a very hazy, indistinct area as to what was the cause, who was responsible, and what can be done to fix or correct a problem, fault defect or error. The user can make choices at the highest level, and basically make wrong choices as to how to solve a data or statistical problem.

A clear example here is the R (or S or S-Plus or IML) language for solving statistical problems. The high level structure of the software allows for very simple command structures, but is heavily dependent on a complex option and default structure (settings, preset commands, etc.). If the user relies on the built-in defaults and gets a wrong answer (or error, or fault), there may or may not be a fault in the R structure.

If however the software does not allow changes in the defaults (for example Excel), then the fault lies with the software. If the user does not want to make the effort to change the defaults, when they can be changed, then this is a problem, not a fault or error. If there are alternate methods for a solution of a problem (for example a test for two means), and the software does not clearly identify the method it uses, then this is a fault. If the software does not also provide solutions by other accepted methods, then this is a problem. An incorrect output number can be a problem, fault, defect or error, depending on the size of the difference, whether the method is exact or an approximation and just what the use is of the number.

At the seventh level, the faults and errors come from the fact that the computer is not a mathematical object, but only approximates it. This is discussed in section 3 on accuracy.

Monahan (2004) comments that students are annoyed at displays that show 3.9999999 but accept 4.0 without question. He also observes that students live in a state of denial as to the possibility of the computer making an error. (My experience with trying to get comments from software reviewers and editors on the issue of software errors indicates this state of denial exists throughout academia.) This is perhaps one reason why so little appears in publications on faults and errors in the popular, large software packages.

WHAT IS A PROBLEM, DEFECT, FAULT OR ERROR?

To clarify these terms with a view towards the provider (developer) of the software, I am going to paraphrase some material from Aglassinger (2000). This is important, because

there is a difference in what these terms mean to the Excel user and the Microsoft software developer. These terms have a certain meaning to a user of Excel. They have a different meaning to a computer scientist (or programmer) who has to make changes to the computer programs (routines) that the user sees as an Excel Worksheet with numbers and text.

PROBLEMS:

“Generally, a *problem* is a subjectively unsatisfactory state”. [For this discussion this usually means that the user won't get the desired output from the program.] “Independent of the actual cause of the problem, the user is unlikely to obtain new versions of the program or recommend it in his circle of acquaintances. This in turn causes a problem for the programmer.”

“There are four ways for the developer/programmer to avoid this situation: first, he can create a program that does not cause the problem to the user. Second, he can create one that helps the user to find out the cause for the problem and remove it. Third, the programmer can refocus on advertising and hope that the user will use the program anyway. And forth, he can use certain techniques to make the user so dependent on the program that he will not consider any alternatives”.

DEFECTS AND ERRORS:

“Problems come from *defects*, meaning a non-fulfillment of a requirement. There are in reality, two sets of requirements. One that the user establishes as an expectation of the program (requirement set 1) and the other, the formal design requirements established by the developer (requirement set 2)”. Many of the defects in Excel claimed by the statistics community, really are differences between sets 1 and 2. An inherent problem is that set 1 comes after set 2 has been established as a baseline for the software.

Formal design requirements for the statistics portion of Excel, originated from some textbooks from the 1970's. The manual for version 4 lists the books. The evidence indicates that it was changed for version 5. The manual for version 5 lists Abramowitz (1972), Box (1978), Devore (1991), McCall (1990), Press (1988) and Sokal (1981) as sources. The evidence is that much of this older structure was retained even up to version 12. There have been major changes in statistics since the 1980's¹, and also, a growing base of Excel users that have training in these newer methods. The community of users from 1999 to the present has a much higher expectation of Excel than what Microsoft was willing to provide.

Secondly, there are some major defects in the “cookbooks” (Press 1988), which were probably carried over into Excel. These C routines (with the defects) probably became a part of the formal design baseline.

¹ An expansion in methods. Extensive time-series analysis tools, extended multivariate analysis, non-linear regression, logistics regression, sampling problems (planning, bootstrap, jackknife, multivariate imputation), experimental designs and ability to reduce data from the design, ability to do simulations, apply Bayesian inferences, Markov Chain Monte Carlo methods, data smoothing (curve fitting and smoothing within regions), quality control, reliability, survival analysis and optimization.

There is no evidence that Microsoft set up a review board that represented the set 1 community. Consequently, what Microsoft perceived as the set 1 community was that of the textbooks listed above.

“Defects simply exist. However, most defects can be removed by performing changes on the item that caused them. An error is a model of the underlying defect that makes it possible for the program to detect and handle it.” (Aglassinger 2000). Errors then are a subclass of defects. There is a distinction between defects and errors here in the programming effort. Note that now ‘error’ has different meanings to the programmer and user. We then have:

- (a) *The programmer’s view that errors are something that can be modeled and handled by internal error handling routines (“Try”, “Catch” and “Finally” blocks in *.net programs).* This is a type A error. In terms of the program, somewhere in the “stack” of “calls”, an “exception is thrown”, and if an error handling routine exists that reacts to the “exception” and produces a message that allows the user to correct an input and continue with the computations, is a type 1 error. The appearance in a cell of error messages such as #DIV/0, #N/A, #NAME?, #NULL!, #NUM!, #REF! and #VALUE! are type 1 errors. This includes a class of “thrown exceptions”, where the algorithm cannot find a solution, or input parameter values exceed some present limits, or that the routine fails to converge.
- (b) Given that a user’s view is, ‘an output from a statistical distribution has errors (Knüsel 1998)’, the programmer’s viewpoint here would be *that all internal errors are correctly dealt with by error handling routines, and therefore it is something that can’t be fixed.* This is what I call a type B error. Type B errors are hidden errors. The programmer is unaware of the error, and the *.xla files are password blocked, so that the error is hidden “to the world”. Sometimes they are indicated when the output values just seem to be wrong. Sometimes a different algorithm or different approach can eliminate type B errors.
- (c) *Errors in the input data put in by the user that produces unsatisfactory outputs.* This is not due to any internal program defects. The user finds this unsatisfactory, since his error is not detected, and ends up giving incorrect outputs. This is what I call a type C hidden error. There are a lot of criticisms of Excel because these users expect correct outputs from incorrect inputs. KBA 829252 was written apparently to address a lot of criticism about input errors. Imputation is a major separate statistical concern, and there are no universally accepted methods to deal with missing data. The user’s community (set 1) expects Excel to be able to “dodge around blanks”, and still give correct values. The problem is basic to Excel, in that there is no variant that has the “missing data” designation. See part 3 on the variant structure.
- (d) *An error that causes the program to “stop”, “lock-up” or “crash” is a type D error.* Type D errors are the result of a thrown exception where the particular called segment does not have an error handling routine for it.

“An error is caused by a *defect*. The problem would even be there if there wasn't any error. A program can happily crash or lose data without the programmer wasting a single thought on any possible defect at all. Putting it that way, errors are a model to deal with

defects. In most cases, an error still imposes a *problem* for the user, as it interrupts the normal control flow of the program or produces an erroneous output. Therefore it is considered ‘subjective unsatisfactory’’. (After Aglassinger (2000))

FAULTS:

Problems also come from faults. In this article, I use the term faults to be those requirements from set 1 that are not met in Excel. Some of the requirements from set 1 are “opinions”, which make it difficult to definitely state that a fault exists. Citations on faults in Excel come mainly from statistics teachers and researchers, who are familiar with more extensive software packages, and expect the same of Excel.

One paper that severely criticizes Excel is that by Cryer (2000). Cryer’s comments are frequently repeated and used to support a position to not use Excel at all for any statistical application. Note H is a summary of Cryer’s (2000) criticisms and an assessment of their validity.

Another term that I use is the term “Buttons” (i.e. push buttons). These are distinct separate menu (tools) selections, tab selections and control inputs A button is “pushed” each time the mouse pointer is on a selection, and is “clicked. A fault in Excel according to Cryer (2000) is that it doesn’t have the right buttons or it takes too many buttons to create acceptable graphics. According to Cryer (2000), this is a fault of Excel.

The correct view is that this is a problem. There is a small fraction of instructors and professors in academia who do not want to build usable general-purpose worksheets and to pass them out to their students to help them use Excel in solving statistical problems. The issue is one of making the effort to change defaults.

The Analysis Tool Pack – Data Analysis – Histogram is a “button”. It brings up an input form and results in a table and a highly compressed chart. The method of entering data, the options and the actual chart is pretty bad. This is discussed in section 17.

R/S and S-Plus are popular because they are one-button packages.

REPORTED GENERAL PROBLEMS WITH EXCEL:

The following is a summary table of problems with Excel in general, giving the application, the problem and a fix or workaround. Note that there has been a marked reduction in the reporting of Excel faults and errors since 2000. Consequently there are very few reported errors with Excel 2003 and 2007.

The reported problems cover a wide range of Excel usage, and they are in most cases found in both statistics applications and in all other applications. The problems reported here are not specific to a limited set of statistical graphics, statistical functions and statistical data analysis routines. Such problems as limited graphics, inadequate documentation and operational problems are important to statistical users, but are also important to other applications and users. Statistical users now use the entire Excel capabilities, including VBA written programs and macros and all the expanded capabilities of the office 2003/2007 suite, including worksheet collaboration and user sharing.

Table 2-1: General Excel Problems

Application or Function	Problem	Source
Help Screens	Incorrect, incomplete or no response from help windows	Note F
Statistical Analysis Documentation	Excel does not provide a log or other record to track what you have done	Goldwater 1997
Faulty computations when missing data occurs	Missing data results in errors in outputs. Excel does not work around missing data. Inconsistent function responses to missing data.	Cryer 2000, Goldwater 1997
Application or Function	Problem	Source
Data Input	Missing data on inputs to functions	Simon 2000, KB829252
Cell Naming	Conflict between range names and certain VBA function names.	Volpi 2006
Range Inputs	Range inputs to functions are not updated when cells are deleted within the range.	RSS 1996
Statistical Functions and Routines	Limited to a rather small set	Cryer 2000
Different Language Version	Really bad translations of English terms/words/usage to the specific language version of Excel	Volpi (2006c)
Automatic Default Changes	When Excel encounters a cell containing what it considers to be a non-standard date or number format, it will automatically “correct” it.	De Levie (2006)
Equation Precedence Differences	A fundamental fault in the way that equations are entered into cells.	Berger (2006 and 2007)

HELP:**OVERVIEW:**

Each Microsoft Excel worksheet function is provided with a *help-file* that indicates the purpose of the function, and includes descriptions of the inputs, outputs and optional arguments required by the routine. This information is usually sufficient to enable a user to make effective use of the function. In addition there is the internet connection to additional help information. Microsoft’s extensive knowledge base articles (KBA’s) are internet accessible, that may have an article on the function, and on problems encountered in using it. Microsoft has a good search engine for finding information in KBA’s. Microsoft keeps these in a review cycle, to ensure currency.

The help screens for Excel 2000 are directly from the Excel Function Reference manuals (Microsoft 1992a) with some new material. The help screens for Excel 2003 and 2007 are

new and in some cases contain new material. There are over 400 functions and routines, in Excel, which have come from many different programmers, groups and organizations. Obviously HELP will not be consistent and both errors and omissions can be expected.

For Excel 2003, the help section and its windows were extensively changed. Most of the text was revised so that the comments about the faults and errors in the Excel 2000 help screens are moot. The issue then of Excel 2000 help faults and errors has no relevance, because nothing can be done about them, except to point them out.

The help for Excel 2007 comes from an elementary file set that is loaded when Excel 2007 is installed, and a more extensive help from the Internet (the linkage is transparent), Help in Excel 2003 and 2007 was not extensively examined for faults and errors.

FIXES AND COMMENTS:

- (a) Help rarely provides full information about the numerical algorithms the function uses. Consequently a black-box approach to testing each function is appropriate” (CISE 27/99 Para 2.1). For many functions, the equations used in the algorithms are shown and the algorithm can be inferred. In many other cases, the method is not indicated, and the function becomes truly an unknown black box.
- (b) Specific errors and faults in Help for Excel 2000 are discussed in Note F. Recommended corrections are also given. There are obviously other errors in Help, but the ones given were the only ones reported in the literature. Excel 2003 has an entirely different help structure. Note F does not address problems and errors in Excel 2003.
- (c) Microsoft has gone to “pop-stat” for Excel 2003 and 2007. “Pop-stat” is the conversion of complicated mathematical and statistical concepts to a simplified view (text) that eliminates the subtleties in the mathematics involved in the concepts. This tends to lead a user to misapply functions or use the wrong functions in analysis. “Pop-stat” also occurs in textbooks in their attempt to explain difficult mathematical concepts. There is a lot in the Excel 2003 and 2007 help text that statistics teachers will find misleading.
- (d) Excel internal structure is complicated. The Excel VBA object model takes over 350 pages to list. Most of this is used by developers in building commercial Excel applications. No one book covers the territory or is fully complete.
- (e) Use Help screens. Going into help with the name of the function or routine is the best method. Help is not as extensive as a normal book index, and can be very frustrating if you don’t know the right words. The Help in Excel 2007 also lists alternate or similar functions or routines.
- (f) Those of us who work with Microsoft products recognize the inadequacy of the help screens at times. Fortunately there are 16 newsgroups on the Internet devoted to Excel (see Note D, “Excel Help From The Internet”) that one can post a question on. Given the huge audience out there of users and developers, one is likely to get a helpful answer from someone.

STATISTICAL ANALYSIS DOCUMENTATION:

OVERVIEW:

One of the reported shortcomings of Excel was a lack of analysis documentation. This was improved for Excel 2007. In office 2007, by clicking on the Office Button → Prepare → Properties, a large ribbon appears with open boxes to put in information:

Author:

Title:

Subject:

Keywords:

Category:

Status:

Comments (a large box):

This information is retained with the workbook

For Excel version 2000 and 2003, consider the first worksheet as an information table with the Excel 2007 document information put in cells. Row and column sizes can be changed to get the information into each cell, without “screwing” up the “data” worksheets structures.

Any statistical analysis documentation requires planning for each analysis, establishing a structure for inputs and outputs and a traceable/repeatable process of the analysis. Two common methods to fix the analysis documentation is, 1, to code the file as read only when finished and/or 2, make printouts for historical documentation. Printouts are probably the easiest way to document the analysis.

Note I gives a general approach to better documentation, useable for all Excel versions.

MISSING OR DELETED DATA:

OVERVIEW:

The issue on missing data is one problem that is frequently reported. Missing data, and ways to work around missing data from a statistical standpoint is really a very complicated issue (see Little 1987). Incompleteness is a troubling feature of many datasets, and assumptions on how missing data is to be treated or assumptions on how new values are to be generated (called imputation), can make very significant changes in the statistical conclusions. The simplest solution to the problem is to delete incomplete subsets within the data set.

Missing data will at times give erroneous outputs. KBA 829252 addresses the missing data problem in the Analysis Tool Pak t-Test that is named the Paired Two Sample for Means tool. Here the routine will give wrong results if a blank occurs in the range. This problem was not fixed in Excel 2003 and remains a problem.

The way data (in cells/rows) is deleted affects the inputs to functions that have as inputs, a range that included the deleted value.

(a) Each cell in the worksheet is linked to data, where the data is in the form of a variant. Eighteen data types are recognized, including empty (un-initialized), null (no valid data) and 16 others representing valid data.

(b) If the number in one cell is deleted by the use of the “Delete” key, the cell becomes blank (null). The (blank) cell is treated as missing data, and skipped in the numerical calculations within the function. If the deleted cell is one of a data pair (i.e. t tests and regression (LINEST)), then the function output will be incorrect. This is a very frequent source of incorrect outputs.

(c) If the number in one cell is deleted by the use of the menu sequence, Edit → Delete → Shift Cells Left/Up, then the ranges that include this cell are automatically changed in all functions. This may create blanks elsewhere within the data set and the Para. (b) condition occurs.

(d) If a row or column is selected, and the “Delete” key used, then the entire row/column becomes blank. The range inputs to all functions are not changed, and the Para. (b) condition occurs.

(e) If a row or column is selected, and the menu sequence Edit → Delete → Entire Row/Column used, then the ranges in all functions are automatically changed. Row numbers and column letters are changed to preserve the numerical row sequence and column alphabetic sequence. This may have unintended consequences for some functions outside of the worksheet region that is the focus of this operation.

One of the barriers to handling missing data is the fact that when a number is expected (in addition, subtraction, multiplication or division) and the expected number is “unknown value”, there is no way to include “unknown value”. The IEEE conventions do not identify a specific entry of “unknown value”.

Users of other commercial software, find that their software “deals” with “unknown values”, and assume (i.e. a set 1 requirement) that Excel does it automatically too. A fault is claimed when it doesn't.

FIXES AND COMMENTS:

- (a) From a simplified standpoint, do not use Excel if you have missing data and you will not or cannot delete incomplete subsets to obtain non-empty arrays or lists. **Any range input should not have any missing data or blanks.** If a range for input to a function has a blank, the output of the function should be taken as possibly being in error, even if the routine is known to ignore blanks. The reason being that these routines may give inconsistent outputs (Goldwater 1997). Imputation to fill in blanks using maximum likelihood and Gibbs sampling, for example is beyond Excel's capability. Simple imputation of averages will bias the outcomes and may result in false conclusions.
- (b) For large data sets, the =COUNTBLANK(range) function in a cell outside of the range can help in detecting blank cells within the range.
- (c) Preferably use the Edit → Delete sequence.

RANGE AND FUNCTION NAMING REFERENCE ERRORS

Volpi (2006b) reports on a user error problem he calls "the conflict between the space naming of macros and cells". It has to do with a user first entering in a cell the standard "=" followed by a function name and valid arguments. Following this, when the user selects that specific cell and gives that specific cell a name (Menu → Insert → Name → Define), the Excel box (Define Name) shows that the entered name is on the "Names in Workbook" list, and that the selected cell is in the "Refers to" line at the bottom.

The problem arises when the defined name is the same as the function name. When the assigned name is entered into another cell, a #REF" error shows up in the reference cell.²

The "User's Guide for Microsoft Excel", version 5 describes how cell ranges can be given names. It recommends the use of the name box and range names with underlines between words to create one-word range names. There is no mention of conflicts between function names and range names. The Help sequence in Excel 2003 ("About labels and names in formula's" → "Using defined names to represent cells, constants or formulas"), is helpful, but does not say anything about conflicts with cell names being identical with the cell = equation function name.

The problem is solved by changing the cell name (See Help, "Change or delete a defined name") or by changing the VBA function name. Volpi (2006b) solved the problem by, "if we open the function wizard we note that the function name is changed. Now the complete function name is 'modulo1.CBRT'³, meaning that Excel does not differentiate the "Name" assigned to a cell and the name assigned to a user defined function. If we re-insert the function with this long name =modulo1.CBRT(x), the thing goes OK. And the bug is also more hidden if we try to add the function CBRT in an add-in "*.xla."

LIMITED STATISTICAL FUNCTIONS AND ROUTINES

From a statistician's viewpoint, Excel has a very limited set of tools for solving real world data analysis problems. This deficiency is frequently cited in the literature, as the main reason to not use Excel. There are add-ns that provide some additional capabilities, but Excel with the add-ins, still is limited. One has to go to the big packages such as SPSS for large statistical problems. Section 19 discusses some of the add-ins.

DIFFERENT LANGUAGE VERSIONS OF EXCEL

Microsoft issues Excel in different language versions.

According to Volpi (2006c), "In the rest of the world, the Excel local versions are a source of lots mess. I will explain better.

"We all agree that translation of the documentation, help-on-line, tutorial, etc. is always a great achievement. In my opinion it is one of the most important points explaining the large diffusion of Excel in the 1985-95 years.

² Volpi's illustration is with the Italian version of Excel. Repeating his illustration in the English version of Excel (2003) the "#REF" error immediately shows in the function cell after executing the naming sequence.

³ The VBA function is in the VBA project directory under "Module1", which occurs automatically when the VBA function is written or imported.

“The menu translation, I dare say, is not so important: nobody is hurt to read "file" on the menu even if the word is not in our dictionary. But somebody may smile reading the menu "window" translated in "finestra" (Italian) or "ventana" (Spanish).

“The function translation, on the contrary, is absolutely a bad thing. It is true that the =MMULT is a good, short, name for matrix multiplication. But in the Italian version the same function appears as =MATR.PRODOTTO, and even more of a long absurd name in the Dutch Version! (Several Dutch friends of mine complain about it).

“We could add also the local translation to the long "Excel complains list".

The differences become a real “pain” when a reviewer has to go over Excel worksheets originated in one language, and try and deduce what was causing problems. The *.xls file from Volpi comes with the entire Italian wording, which is acceptable. However the message boxes, error inserts, etc. are also Italian, which makes it hard to find errors and faults, of the type discussed at the very beginning of this section.

EQUATION PRECEDENCE ERRORS

One of the undefined critical problems in regard to equation accuracy has to do with how the equation is formed in Excel. This is the precedence problem.

Berger (2006 and 2007) points out a fundamental fault in Excel is in the way that equations are entered into cells. The fault is in the way that negation is expressed as a minus sign. “To see this in Excel, enter = -3^2 into a cell. The result Excel reports is +9 because Excel interprets this as (-3)^2. On the other hand, = 0 - 3^2 results in -9, because now the "-" sign is interpreted as subtraction (lower order of precedence than exponentiation) rather than negation.”

Asseburg (2009) wrote “One thing that I've noticed (and also found discussed on some internet pages and in a KB article by Microsoft) is the precedence of unary minus operator. Microsoft claims that this is not a "bug" in Excel because they did describe in their help that they are not following the usual mathematical precedence rules.”

In KBA 132686, Microsoft stated, “In Microsoft Excel, when you use a minus sign (-) as a negation operator (for example -1) in a formula, the negation operator has higher precedence than a binary operator. This order of precedence may mean that a formula returns a positive value when you expect it to return a negative value. For example, the formula ‘=-2^2’ is evaluated as ‘(-2)^2’. The minus sign is evaluated as a negation operator. The formula returns a positive value, 4.”

“Microsoft Excel uses an order of calculation to evaluate operators in formulas. The order of evaluation of operators dictates that a minus sign (-) used as a negation operator (such as -1) is evaluated before all other operators. Because of this order, the formula ‘=-1^2’ represents the value -1 squared, and returns the value 1, a positive value.”

“That has been the standard method for evaluating formulas since the first version of Microsoft Excel. “

It is highly unlikely that any Excel user will have ever run across or read this KBA and as a result modified an equation to establish the order of precedence's by use of parentheses.⁴

The philosophical issue is that Microsoft should base Excel usage on the way that users were taught about arithmetic and the forming of equations in grade school and in high school. A user should not have to make an effort to read an obscure KBA first before entering simple equations into Excel cells.

Section XX further discusses this problem, which may be the cause of some of the errors in Solver solutions.

MIXED TYPE B PROBLEMS AND FAULTS

THE COMPLEXITY OF EXCEL

A lot of these problems come from the complexity of Excel, and the changes in appearance, menus, capabilities and "feel" from version to version. Some of the practices a user is comfortable from version 5, lead to errors and faults in version 12.

Excel is a large complicated program with many features, functions and routines. To effectively use Excel, one should have reference books such as Ivens and Carlberg (1999), and know how to get information from it. At a lower level, one needs to have gone through an introduction to Excel, such as that found in the first part of the Excel manuals that are required as part of a statistics course that includes Excel (Dretzke 2003 and Dretzke and Heilman 1998). Some of the reported problems with Excel were the result of an inadequate knowledge of Excel on the part of the user.

Faults and errors occur in user written VBA functions and subroutines. VBA allows one to automate a lot of repetitious operations and to do calculations and change displays, appearance and to link with other workbooks and data sources. VBA cannot be written without learning the intricate fundamentals of the very complex object and linkage structure of VBA⁵.

SINGLE-CELL OR ARRAY FORMULA ENTRY

OVERVIEW:

There are two basic ways to exit a cell containing an equation or function:

1. The single cell output and exit (Push the Enter key or select the entry line menu green check mark)

⁴ Making inquiries into Microsoft's Products inquiry website using the term "order of precedence" produces no response, useful information or direction to this KBA.

⁵ Microsoft had published for Excel version 5, the books, "Visual Basic, Getting Started", "Visual Basic, User's Guide", "Visual Basic, Component Tools Guide" and "Visual Basic, Programmer's Guide". These were very helpful in learning about a very complicated system of commands. For current versions the commercial books such as, "Excel 2000 VBA" is needed, because of the very complicated object model that is VBA. It is this complexity that makes it difficult for users to go into VBA to do computations, graphics and displays.

2. Multiple cell output and exit. This is required when the “equation or function” generates multiple outputs. This is called array-formula entry. This **REQUIRES** the following sequence:

Use the mouse to select an output range of cells that will hold ALL the output. This requires some planning to determine the range of output cells the formula will fill. Error on the larger size, since the unused cells will end up with N/A symbol. The range is rectangular.

Move the pointer to the upper left cell of the output range. It should be flashing.

Put into this cell the formula with the beginning equal sign, and with the inputs having the correct external data input ranges entered.

Do the keyboard sequence CTRL-SHIFT-ENTER.

RSS (1996) points out that array functions are not supported properly. This was in earlier versions of Excel, which has been fixed for Excel 2003.

Note E identifies those special functions that require array-formula entry.

Another source of array-formulas is when a complex formula is built up from the use of SUMPRODUCT or SUM, or when simple combinations of multiplications (and sums) involves a range of cells as a variable rather than a single cell. For example the formula $=B\$1:\$F\$1*B\$2:\$F\2 placed in cell D4, requires the pre-selection of the range D4:H4, formula entry in cell D4, and CTRL-SHIFT-ENTER to give the correct output (Ivens 1999).

FIXES AND COMMENTS:

- (a) Ignoring the requirements of array-formula entry is a very common problem. It frequently shows up on the news lists. LINEST requires array formula entry. You have to select a column range space equal to the number of variables plus one, and a row range space of 5.
- (b) Don't rely on the function wizard. Array functions have to be handled in peculiar ways. Not easy to put array functions in one cell. The DETERMIN function is an exception. LINEST is an array function, and Microsoft addresses this array entry problem with LINEST in one of their knowledge base articles.

DATA INPUT PROBLEMS:

Spreadsheet setup and data entry may at times create hidden errors. These are not detected, and carry over into faulty or wrong outputs. This has to do with the way the external/internal world is described by using symbols, language and numbers. It also occurs from the way that Excel as a spreadsheet works with data. The Help windows on “Arguments”, “Numbers”, “Text”, “Logical values”, “Error values”, “References”, “Arrays” and “Converting Data Types” give information on how Excel handles inputs.

Note G, “Data Input Errors”, discusses this problem and the errors from using numbers to represent symbolic or categorical data.

Excel has some peculiarities in data entry that are different for each version. In Excel 2003 and 2007 one occasionally gets locked in text-recognition that can't be changed. In this case, select the cell, go to Edit and select Clear and then select All. This will clear the cell of everything. Now enter the number.

Hidden errors can occur when data are entered in Excel 2000. The contents of each cell are stored as a variant type (128 bits). Within the variant is a code telling Excel which of the many types, the entered data is. For example with the *General* format on each cell (the default), you enter the following, followed by punching the Enter key.

A number (digits). The variant type is set as double.

A number with digits and accidental spaces before and after.. The variant type is set as double.

A number with beginning quotes. The variant type is set as text.

A number with an accidental (not numerical) character. The variant type is set as text. You recognize the error and re-enter the number. The variant type for that cell remains as text. Formatting the cells (range) as number will not convert the number type back to variant. You will have to select Edit, select Clear then select All to remove this hang-up.

Sometimes Excel 2000 logic will change the variant type to double and sometimes not. If the cell is initially empty and the cell default format is "General", then Excel correctly sets the data type. If the cell had an error and is being changed, Excel may not correctly reset the type. If not changed, the hidden error is now in the cell. If later on, this list is a range, and the range is input to STDEV, the output from STDEV will be in error because the cell was skipped in the calculation, and there was no evidence that this error had occurred.

This is an example of what De Levie (2006) calls the automatic default change error. As he says, "If you use Excel for genomic research, and want to represent, e.g., the tumor suppressor called *Deleted in Esophageal Cancer 1*, abbreviated as DEC1, Excel will change DEC1 into 1-Dec automatically, and without warning. The standard abbreviations for septin proteins will likewise be changed from, e.g., Sept2 to 2-Sep, and, even more troublesome, Excel will change the RIKEN clone identifier 2310009E13 into 2.310009E+19 or even 2.31E+19! These problems are especially pernicious when large data files containing such abbreviations are imported into Excel. A possible work-around is to place a space or an apostrophe in front of all such names (before they are imported into Excel) or to pre-label the receiving columns as text, so that their contents will be considered as text strings."

Microsoft can't be faulted for the occurrence of these errors, since the computer logic that does this is a reasonable approach to the ambiguities that humans use in language. However Microsoft could make it easier to change rows, columns and ranges completely to number, date or text types, regardless of what had been set based on the input. If the computer interprets the entry as a date type, the original entry is lost, and cannot be reconstructed from the date.

Although you can change the format of the cell to text, this does not change the data type

Excel 2003 and 2007 have a tendency in a cell to sometimes lock into the text type and cannot be changed back to numbers by imputing numbers. One has to go to “Edit → Clear → All” for that cell to allow it to be reset. The lock-ups also tend to occur when dual keys are accidentally pressed.

Errors in a sort or even loss of data will occur if in a column of numbers, one or more numbers are internally represented as text. KBA 214282 describes the problems with sorting, when numbers and text in one column is sorted. In ASCII (text mode), a zero has more ones bits than 1, and results in some odd sort sequences.

Another way to deal with missing and data errors is to set up data validation for a range of cells from the menu bar. Data Validation however only works for individual cell entry.

THE CELL EQUATION JAMMING PROBLEM

The requirement that any cell equation be “one line” forces considerable complexity and artfulness to nest the calculations as a single line (or to remain in one cell). The nesting of functions leads to many invisible errors. Separating punctuation is very small, and results in error prone strings. Subsequent users can tinker with these, creating hidden errors.

One sometimes encounters the “3 inch rule”. This is a working limit on the length of any cell equation.

All of the above are user errors and faults. In a sense, Microsoft can’t be faulted when these occur, but Microsoft can be faulted by not allowing the use of multi-lines and blocking (such as that used in VBA), to simplify the structure. The forcing to a single nested line results in a high proportion of errors in cell equations. There are those Excel experts who delight in being able to build long dense cell equations of characters that are almost impossible to understand. The expansion for Excel 2007 on the limits of the number of characters in a cell from 256 to 8000 allows for extremely dense If equation structures in Excel 2007. If the equation cell is not locked, an inadvertent insertion/deletion can make the whole worksheet unusable.

ARulesXL (<http://www.arulesxl.com/>) is an add-in that allows considerable simplification of single cell IF structures. Panko (2005) comments on the errors in these long IF structures, and on the ability to find errors after they are written.