

I. INTRODUCTION	1
TEACHING EXCEL IN STATISTICS COURSES.....	1
THE VIEW (CA 2000-2006).....	1
USE AND APPLICATIONS OF EXCEL.....	4
BUSINESs	4
TESTING, MEASUREMENT, EXPERIMENTS.....	5
GENERAL DATA ANALYSIS IN EXCEL.....	6
OTHER SOFTWARE FOR STATISTICAL COMPUTATIONS.....	6
EXCEL ERRORS, FAULTS AND FIXES	7
EARLIER EXCEL VERSIONS	7
COMPATABILITY	7
USE OF EXCEL 2007 IN TEACHING AND LABORATORY SITUATIONS.....	8
EVALUATION AND REPORTING OF ERRORS AND FAULTS.....	9
THE VIEW (2009 AND BEYOND)	9
Issue 1:	10
Issue 2:	10
Issues 2 and 3:.....	11
Issue 4:	11

I. INTRODUCTION

TEACHING EXCEL IN STATISTICS COURSES

THE VIEW (CA 2000-2006)

There are schools that teach an introductory course in statistics using Excel as a computing tool.¹

The issue here is that spreadsheets are universally used in business, government, education, research, manufacturing and just about all other sectors. Spreadsheets are an essential backbone in every aspect of what people do. Typically they are employed to:

1. Gather, create and manage data of all kinds; numbers, text, symbols, observations, surveys, money, financial, accounting, designs, etc.
2. To create models and make calculations.
3. To disseminate and share information in a basic form across a wide expanse of users and contributors in a completely connected world.

¹ The actual practice of teaching how to use Excel (within the course) has been declining since about 2001. This is a separate issue.

Spreadsheets have become indispensable tools for getting the informational work done. They are empowering tools that are expressive and apparently simple, yet underneath very complex. Text and numbers can be intermingled. They can be subservient in that they facilitate peer-to-peer sharing, non-technical people can do analysis and share the data and results. They facilitate back-channel, behind the scenes communications.

They also create enormous problems with errors in data entry, errors in equations, misuse of data sections, and incorrect use of functions. (This is discussed in section 2.)

Business, Engineering, Psychology, Accounting and other schools in colleges and universities find they have to teach the intricacies of using spreadsheets such as Excel, because of its universal use in all sectors of the world. Statistics in a broad sense involves every one of the three areas that Raden (2006) points out. Therefore, why not teach statistics in a manner that involves spreadsheets? The other side of the coin is, if Excel has to be taught, why not include teaching about the use of the statistical functions and routines in Excel?

The University of Reading (SSC) has emphasized the use of Excel in statistics and the use of spreadsheets for entering and tabulating data. Arshan (2007a) has a web site that covers all the essential Excel statistical capabilities with respect to business applications, as part of an MBA degree program. Arshan (2007b and 2007c) has other sites that show extensive use of spreadsheets and cell equations to handle business decisions and other problems.

The intent of teaching the use of Excel is to produce graduates who can use the powerful spreadsheet capabilities of Excel, have some understanding of statistical methods and can do quantitative reasoning with statistics. Some other schools combine Excel with other software programs such as Minitab. Levine (1999) and Pelosi (2000) are some of the more frequently used business statistics textbooks that incorporate Excel.

Levine and Fan (2000) say “The strongest aspect of the book (“Practical Statistics by Example Using Microsoft Excel”, Sincich, Levine and Stephan, Prentice Hall, 1999) is the introduction and incorporation of Excel for doing and learning statistics. The authors rightfully argue in the preface that ‘today an increasing number of individuals use spreadsheet application as the means to retrieve and analyze directly the data they need. Employers now are beginning to desire, if not demand, that their college-educated, entry-level employees have more than just a cursory awareness of spreadsheet applications. Most students are familiar with Excel and/or have easy access to it on a personal computer...’ They also point out that there are real dollar and time advantages to universities when Excel is used as part of the course. It allows students to do homework on personal computers², reduces the load on school computer labs and saves the cost of the expensive commercial licenses for use on each pc.

² There are legal problems here. The license to use the software may prohibit its use on student (home) computers. The license may be so restrictive that the student cannot use it outside of the school’s computer laboratory. If the student buys the academic version, this then gives unlimited academic usage. Microsoft has academic versions of its software. Most Excel add-ins that come with books are so restrictive so as to be totally useless to the student beyond the course.

De Levie (2005) finds that Excel is useful in science and engineering laboratory and application areas. “Excel is a powerful spreadsheet. Even though it was developed primarily for business applications, it contains many mathematical functions, and its ease of use and wide distribution make it a very powerful tool for scientists and engineers.” Some of the textbooks are Billo (2001), Block (2000), de Levie (2001), de Levie (2004), Gotfried (2000), Liengme (2002) and Orvis (1996). The Visual Basic for Applications (VBA) feature with the ability to program specific reductions and analysis is important in engineering and scientific areas. The main statistical usage here is equation fitting and regression. Graphic display of data and the display of data fits to mathematical relationships is also important.

The use of spreadsheets as a means of explaining subtle ideas by doing simulations is another valuable teaching method (Horgan 1999). See the great spreadsheet by Jacob Eisner (Eisner 2007) to teach the forward-backward algorithm to solve a probability problem. Some of the more difficult ideas in experimentation, sampling, variability and power can be demonstrated.³

This paper refers to the following textbooks, which all incorporate the use of Excel for solving statistical problems. This is just a sample of many other textbooks that incorporate Excel.

1. Larson and Farber 2003
2. Levine, Berenson and Stephen 1999
3. Lind, Marchal and Mason 2002
4. Moore and McCabe 2003
5. Pelosi and Sandifer 2000
6. Triola 2001.

Texts 2, 3 and 5 are focused on business and economic applications. Text 4 also shows how to use JMP, Minitab, S-PLUS, SPSS and SAS. I am familiar with the use of 1 in an introductory statistics course.

However including Excel in an introductory statistics course has its own problems. The inclusion of Excel is controversial and is not universally accepted. Some teachers in university statistical departments disparage the use of Excel as a tool for doing statistical calculations. Others have tried it and found severe problems, and have discontinued the practice. Some of their arguments and criticisms are:

- It takes teaching time to teach Excel⁴. Most introductory statistics courses are very time limited to teaching the material in the textbook. Teaching Excel takes

³ The media has shifted now by 2008 to the use of the Internet. Interactive programs are easily accessed by students and they are far more effective here than the static Excel worksheet/graphics..

⁴ Gentle (2004) argues that this training be provided in the “just-in-time” mode and be extended into other courses that use the software. There is a need for University short courses that do this, that should not carry academic credit, but neither should they carry the enormous fees of commercial enterprises. He also observes that “on-the-job”, the ability to do computing is largely a “self-taught just-in-time” activity. Be aware that “self-taught-just-in-time” is “problematic, in that what the student picks up may be very incomplete and full of application errors, leading to bad research.

away from teaching statistics. It is a lot easier to teach problem solving and to test students using hand calculators than with computer software.

- The Excel default graphics do not fit standard statistical data displays, and takes considerable teachers and students time to change them to standard displays. Excel 2007 makes it more difficult to make changes with the complex expanded menu structures. Excel is not a “one-button” statistics package.
- Excel functions and routines do not fully support the subject matter or the problems in the textbook (from 40% to 70% of textbook problems are directly supported). Excel does not support expanded applications in line with contemporary statistics. The ANOVA and regression routines are too primitive and size limited. The data analysis routines are too primitive. Many teachers take the view that Excel is not capable of serious data analysis. Excel cannot be used in advanced classes (McCullough, 2004).
- The field of statistics is always evolving, with new ideas and methods to analyze data. At the introductory statistics level, some of this is being introduced. Excel is essentially a locked-in-time (ca 1990) approach to statistics, and has not introduced any new functions or routines (re: Excel 2007), relative to what has evolved since then.
- Uncertainty about the reported errors, faults and inaccuracies in Excel. This is a very weak argument, since it is not raised on other commercial software.
- Excel is not usable for classroom quizzes, tests and examinations based on solving problems to test students for comprehension and understanding. Testing is still a paper-and-pencil process.
- Text-books that incorporate the use of Excel normally include either a CD or a time limited password access to the publisher’s website that gives the data sets, files relating to teaching (illustrations, slides, etc.) and may include an Excel Add-in. Publishers charge more for this combination, markedly increasing the cost of a textbook with the Excel features for students. This is a significant unrecoverable expense for the student. I have heard lots of complaints from students on this.

Note A, describes some of the other reported comments on the use of Excel in a teaching situation. Of particular interest is the paper by Peter C. Bell, (Bell 2000) about his course in business statistics using Excel.

USE AND APPLICATIONS OF EXCEL

BUSINESS

Excel is used as a spreadsheet program in businesses and in government. It is used for all kinds of analysis, including financial analysis, economic studies and analysis, problem solving, engineering problem solutions, management problems and day-to-day business operations. Excel is generally available on individual computers in all large corporations, government organizations, non-profits and most small and medium sized businesses. Excel worksheets are basic for running a business.

Microsoft (2003) reported that there were 400 million licensed Office installations worldwide. Some recent correspondence from some experts on EUSPRIG, indicated that currently there are about 440 million Excel users worldwide and about 160 million users of other spreadsheet programs.

In business and in engineering there are a lot of “small, frequent” problems. As Hillmer (1996) says, “First, managers have a greater need for statistical tools in problem solving than they do for statistical inference. A manager’s main use of data analysis is in the context of problems they face every day... Third, a beginning required statistics course for future managers should be sure to teach the tools, which are most likely to be relevant to solving problems. Many of the most relevant tools are relatively simple because experience has shown that many times simple tools are adequate for dealing with the majority of managerial problems.”

Most of these problems can be very nicely investigated using the spreadsheet capabilities in Excel. The integration with financial and other functions can show effects in terms of future cost, future returns, and expected income. It can augment six sigma quality control efforts, identify “outliers”, indicate ways to reduce inventory, reduce investment costs, increase yield, analyze problems and show unforeseen opportunities. The “What-if” type of analysis in Excel is a basic and important tool in business. In six-sigma quality, an outlier is an opportunity to investigate what caused it, not something to reject by “trimming”. Excel has the ability to pre-process data and post-process results, which are decided advantages. The pivot table feature is very extensively used in business.

Also, “Windows is becoming less ubiquitous, 5% of PCs are not Windows, and this percentage is growing. Mac and Linux spreadsheets can be ported to Windows, but not conversely” (McCullough 2004). However in the business world, Visual Basic, Access Data bases and SQL is heavily used in applications. FORTRAN is not used, and C++ used with Linux is not used much for application packages. R is becoming dominant as a broad tool to solve data analysis problems, and it is free. The outputs are not directly transferable to Excel, except by text based copy-and-paste operations.

The ability to merge Excel with other business applications and to use a common programming language (Visual Basic, Visual Studio, etc.) is a distinct advantage. The Active X capabilities among Microsoft programs allow a lot of blending of real world data with tools to “work” the data.

The Statistical Service Center 2000 states, “Excel offers an exiting environment for data manipulation and initial data analysis. Its pivot tables are particularly good for cross-tabulations and summary statistics and provide a powerful tool for basic data analysis. The reliability of more advanced statistical functions and wizards is variable.” They emphasize the use of pivot tables (“Excel’s pivot tables are very powerful and are an area that is better in Excel than in many statistics packages.”) to summarize data and give an extended appendix on how to create and use them.

TESTING, MEASUREMENT, EXPERIMENTS

This is a very broad area involving data collection from instrumentation and from all kinds of physical devices. This includes industrial laboratories, field data collections, university laboratory courses, product development, etc. The essential core here is that

the phenomena can be measured and the measurement involves instrumentation, where the output is electrical voltage.

The voltage, as an analog signal can be measured, and converted to a bit sequence, which in turn can end up as a value in an Excel cell. For example, DATAQ (www.dataq.com) (and others) have low cost analog to digital converters (ATD). The connection from the ATD device to a computer is via an USB link. Software (www.ultimaserial.com) within the computer then can manage the “collection” of data, do computations on it and generate charts. DATAQ also provides software that display ATD outputs. A four channel, 0 to 10 volt input, 14 bit integer output ATD capable of 14400 samples per second runs about \$120. UltimaSerial Software using Active X then puts the data into Excel cells as normal numbers.

The Excel worksheet route gives a very flexible and low cost route to collecting a large amount of data, and fitting models and “concepts” to the data.

GENERAL DATA ANALYSIS IN EXCEL

Also to be considered are many complete statistical analysis programs that work within the Windows environment. Some are free (internet downloads) but most are commercial that have to be paid for. The commercial software packages can usually download data from Excel worksheets, but have their own peculiar outputs that may not appear as worksheet cells. Some of these programs, as add-ins that work entirely within Excel environment are described in section 19. The free POPTOOLS add-in and other add-ins add a lot of useful mathematical and statistical tools to Excel (but they are not tested for accuracy).

The fact still remains, that Excel has a very limited statistical capability.⁵ Where statistical analysis of data is a main job requirement, learning and using one or more of the larger commercial software packages is a must. However where presentations have to be frequently done (i.e. Power Point), the employee must learn how to use Excel, Power Point and the larger packages together to be effective in his job.

OTHER SOFTWARE FOR STATISTICAL COMPUTATIONS

There are other options that can be considered.

Some of the newer editions of the above listed textbooks no longer include Excel as a means of solving problems. They have temporary internet links to publisher websites that contain Java based computational tools that emphasize the mathematics of solution. These sites are only accessible as part of the course, and if the student wants to use them for real problems, he will have to pay for that, just like other commercial software.

Another option is to teach use of one or more of the free statistical software packages that can be downloaded from the internet. Robert Dawson has a web site (Dawson 2007) that lists some of these, and what they will do. Some may be only useful for plotting data.

⁵ Microsoft has not expanded the basic statistical capabilities of Excel since Excel 4.0.

The site (Dawson 2007) does not list all the available software or the other free add-ins to Excel that improve the capabilities of Excel. The software listed, is also “untested” in the sense that the results of running the StRD test data sets through the software have not been done, so that the accuracies are unknown (with the exception of R and DATAPLOT). Most of them are also limited on what statistical problems they will solve (with the exception of R). The other limitation is that the outputs are in a fixed format (number of digits, usually 3 to 5), so that rescaling of the data has to be done in all cases to be sure of even 3 digits (e.g. no allowable exponential notation).

Computations using the internet and Java software do have limits on accuracies. See Kitchen, Drachenberg and Symanzif (2003) and Kahan (2004). The Intel IA-64 instruction set (Cornea-Hagan and Norin (1999)) is the preferred means of calculation, since it retains some of the benefits of the use of the IEEE-754 Long Double format, and has accurate division.

What we also see is a growing use of “R” as a (free) statistical analysis tool. “R” is completely different from Excel. Essentially, if one uses “R”, data and results have to be manually recorded and manually entered into Excel or manually entered in to “R”. There is no computer level interface. Much of the analysis that was done in the past using Excel, has now shifted to “R”.

EXCEL ERRORS, FAULTS AND FIXES

EARLIER EXCEL VERSIONS

Several articles, a lot of emails on the stat lists, and many Internet sites have stated that there are errors, faults and problems with Excel. Some of these allude to errors in the earlier versions (See Note B for a list of the versions and when changes were made). RSS (1996) describes errors in versions 5 and 7 that were fixed in later versions. Many of the errors were in earlier versions, and Microsoft chose to ignore these, putting off critical fixes until the 2003 version. The Microsoft KBA series describes some of these earlier problems with Excel and describes the changes made for the later versions (see Note C for a listing of applicable KBAs).

The statistical and statistically related functions and routines in the three versions, Excel 97, Office 2000 and Office 2002 are essentially identical. They are combined under the designation “Excel-2000”. Version 11.0 (designated “Excel 2003” or “Excel 2004”) made some major changes. Excel-2003 functions and routines that were changed were also tested and evaluated in this paper. For Excel-2007 (Version 12.0) there were no changes to the basic version 11.0 statistical algorithms. Consequently, when the function/routine has not changed from the 2003 version, the reference will be to “Excel 2003 and 2007”. When different, the references will be to the separate “Excel-2003” and “Excel-2007” entries.

COMPATABILITY

The appearance of Excel-2007 (look and feel) and the basic file designators were drastically changed for the 2007 version. The changes were to give it the look and feel of the recent Vista operating system, to provide an expanded data (size) handling capability and to change the internal structure of files from the *.xls and *.xlm file formats to a compressed, more secure format. The Excel 2007 files formats now have the extensions:

.xlam, .xltn, .xlsm, .xlsb, .xml, .xltx., .xlsx or .xlw, all of which are not recognized by prior versions of Excel. The .xlsx extension is the default file extension. There is also a backup file, which can be created with the .xlb extension. These extensions may not appear if you have the folder hide-extensions setting on.

If you open a *.xls file in Excel 2007 you will be in the compatibility mode. If you make any changes using any of the new Excel 2007 features, the compatibility checker intervenes when you try and save. There are many Excel 2007 features that will not work in the compatibility mode.

Excel-2007 files are not backward compatible with any earlier version of Excel, unless they are saved as *.xls files. This loses all the advanced features of 2007. Excel-2007 will work in the Windows XP environment and in the Vista environment. Excel-2003 will work in both Windows XP and in Vista. Excel 2007 charts can be inserted into Word 2003 documents (by copy and save, then back in Word 2003 by paste. But this does not always work.

Word 2007 documents saved as *.docx are not recognized by Word 2003. They have to be saved in the *.doc format to be compatible.

USE OF EXCEL 2007 IN TEACHING AND LABORATORY SITUATIONS

The look-and-feel for Excel essentially remained the same for all the versions from Excel 97 to Excel 2004. With Excel 2007, there was a major change in the menus, the look-and-feel of worksheets, the accessing of functions and routines, and the appearances of worksheets and charts.

Essentially all the books, textbooks and manuals available for Excel 2003 are now obsolete. Their how-to-instructions do not apply to Excel 2007.

Generating charts for displays is now completely different, with respect to menus, sub-menus, results of selections, and the final appearances of the charts. It takes longer going through complex menus and dealing with single word selections that do not convey what the selection word means. This basically says that all the existing textbooks and CD's with problem data, and publisher add-ins are now obsolete for Excel 2007.

Building charts has completely changed, and their appearances also completely different. There is an increased ability to put in chart-junk', lighting effects, shading, 3D renderings, flashy, distracting figures, silly variations, insertion of icons, visual distractions, etc. This is what the business world wants, the ability to insert effects to obscure, bias or just to add variety to frequent presentations (See Few 2006). This can lead students to become very adept at obscuring what the data shows.

Good, effective, clean charts are difficult to build in Excel 2007. Karl Ove Hufthammer recommends Tufte 2001, Robbins 2004 and Few 2004 and 2006 as good sources for making good statistical charts.. There are other sources for building better business charts such as from juiceanalytics.

The default charts shown in the Data Analysis routines are inadequate. It is quite difficult to start with the general default chart and to create a clean, positive, simple chart showing relationships. This is a real loss for displaying laboratory and statistical results. Colors are

no longer bright and positive, but gray, smudged pastels. The process of cleaning up default charts is too long and involved.

A common use of Excel was to create data plots and data reductions, and then cut-and-paste the chart into a WORD document. This no longer works well when the document is a WORD 2003 document and the cut is from EXCEL 2007. Consequently one has to use the WORD 2007 version of the file to correctly cut-and-paste.

Microsoft has essentially added a lot of complexity that makes it difficult for students to use it as a convenient analytical and display tool.

EVALUATION AND REPORTING OF ERRORS AND FAULTS

It is important that errors, faults and problems in a software package be identified. Altman (2000) points out the importance here of being able to “rely” on the numerical results. He states, “Numerical accuracy can mislead social scientists that are caught by it unawares – so we must pay attention”. A user can be misled by the display of numbers.

McCullough (1998 and 1999) established the current criteria for evaluating statistical software. McCullough and Wilson (1999 and 2000) made important assessments of the capabilities of Excel and the problems encountered in using Excel for statistical calculations.

This paper attempts to consolidate most of the criticisms, reported errors and faults in statistical applications, and to evaluate their claims. The other purpose is to describe workarounds and fixes that overcome these faults and deficiencies in Excel-2000. Problems, faults and errors that still remain in Excel-2003 and Excel-2007 are also discussed. If the problem in Excel-2000 has been fixed in Excel-2003, it will be discussed. If there is no explicit indication of a change in Excel-2003, then it can be assumed that the problem still occurs in Excel-2003 and Excel-2007.

There has been a small but steady stream of articles about Excel 2007 faults and errors in the main-stream professional and academic journals. An index of these has not yet been developed.

THE VIEW (2009 AND BEYOND)

Courses in Statistics have changed over the years with changes in technology, changes in society and changes in demands on the work force. Current guidelines for contemporary statistics courses (in ASA and in Garfield, Hogg, Schau and Whittinghill 2000) now emphasize the need for students to rely on computers using statistical software programs.

The view now is that these computer programs have essentially replaced the need for any researcher to understand the technical and theoretical details of the related mathematics and philosophic concepts. There is a loss then of any focus on statistical thinking during the research planning phases. Brown and Kass (2009) point out this loss as an increasing problem of bad research that gets published.

The essential issues now in post 2009 are:

1. Does a researcher have to have at least a knowledge-of or have some training in statistics at an “Introductory Statistics Course Level? To what extent does his

- algorithms have to be mathematically sound? Is there a real need to deal with random effects and “errors”, or can this be ignored?
2. Is the basic Excel concept of spreadsheets to enter and handle data relevant?
 3. Are the tools that Excel provides usable for analysis?
 4. Are the visual graphics tools provided in Excel useable for presentations and publication?
 5. Applicability: (1) Can a user make an important contribution to a research project or a contemporary data analysis project by using the statistical (and mathematical) functions in Excel 2007? (2) Can a user make an adequate analysis of business data, by only using Excel 2007? (3) Can a user obtain an adequate analysis of contemporary scientific or engineering data by only using Excel 2007? (4) How accurate and useful are the results, including the charts?

COMMENTS ON THESE ISSUES:

Issue 1:

Brown and Kass (2009) comment on the changes that have occurred in contemporary demands for data analysis in all areas of science, engineering, medicine, neurology, business and government.

- a. “The focus in these introductory statistics courses has been to train a student to answer circumscribed methodological questions based on a limited contemplation of context”.
- b. “Much of which is traditionally taught in statistics, is now extraneous to compelling scientific and practical problems that students are interested in solving. Needed statistical education has not been sufficiently accessible. The most innovative and important new techniques in data analysis have come from researchers who would not identify themselves as statisticians.”
- c. “There is a lot of bad research being reported that lacks a strong mathematical and statistical basis”.

Issue 2:

The concept of collecting data and entering it into spreadsheets for evaluation and analysis is a past concept. The commercial software now currently in use, does not use Excel “spreadsheets” for inputs. Some however have separate (limited) routines that will read data from Excel spreadsheets. The widely used (and free) “R” statistical software does not use “spreadsheets” for data inputs. This essentially says that Excel is irrelevant now for any contemporary research.

There is Excel spreadsheet specific software, both commercial and freeware that has been used, and is still available. This is all “user beware” since almost all of it has not been independently tested. It’s easy to write the Visual Basic software and to promote it. There is no incentive to test it for correctness.

Issues 2 and 3:

The basic capability here is still relevant. Spreadsheet adaptable methods such as that in Kline (Kline 2004, “Beyond Significance Testing”) are currently relevant methods. They require effort to layout the data tables, to use the Excel table commands, to setup the equations and relationships between cells and to (possibly) build simple Visual Basic functions. The issue is essentially, “is this a relevant 2009 research tool?”

Issue 3: The statistical functions and routines in Excel are essentially the “tools” of the “past” and represent a view of statistical functions that has remained locked at the traditional “Introductory Statistics Course Level” ca 1990. No new relevant tools have been added in the succeeding Excel versions

Issue 4:

The answer here is no. The section on charts discusses this problem.