The Risks of Using Spreadsheets for Statistical Analysis

Contents:

Introduction

Spreadsheets are widely used for statistical analysis; and while they are incredibly useful tools, they are useful only to a certain point. When used for a task they're not designed to perform, or for a task at or beyond the limit of their capabilities, spreadsheets can actually be dangerous.

This paper presents some points you should consider if you use, or plan to use, a spreadsheet to perform statistical analysis. It also describes an alternative that in many cases will be more suitable.

Why spreadsheets are popular

A spreadsheet is an attractive choice for performing calculations because it's easy to use. Most of us know (or think we know) how to use one. Plus, spreadsheet programs come as a standard desktop computer resource, so they're already available.

A spreadsheet is a wonderful invention and an excellent tool – for certain jobs. All too often, however, spreadsheets are called upon to perform tasks that are beyond their capabilities. It's like the old saying, "If the only tool you have is a hammer, every problem looks like a nail." But some problems are better addressed with a screwdriver, with glue or with a belt buckle.

Moreover, the perception that a spreadsheet is easy to use is, to some extent, an illusion. It is always easy to get an answer out of a spreadsheet – but it's not necessarily easy to get the correct answer.

However, the decision to use something else – an unfamiliar technology or tool – is not always an easy one. When considering an alternative, two questions come to mind: How useful is this tool? And how hard is it to learn?

Spreadsheets can be useful for statistical analysis; but when used for tasks they're not designed to perform, they can actually be dangerous.

The answer to the first question depends on the scale and the complexity of your data analysis. A typical spreadsheet will have a restriction on the number of records it can handle, so if the scale of the job is large, a tool other than a spreadsheet may be very useful.

As for complexity, if you only need a superficial review of your data, a spreadsheet may be a suitable tool. But if you suspect that there is valuable information in your data that isn't immediately obvious, or if you need to perform a detailed analysis or find hidden patterns, a spreadsheet will not give you the functionality you need.

Another factor to consider is the degree of accuracy required. Spreadsheet results can be unreliable, especially on large datasets and/or for complex calculations. If absolute accuracy is required, a spreadsheet may not suffice. Instead, a different, more reliably accurate tool should be considered.

Finally, if the task is simply to analyze a limited quantity of historical data, a spreadsheet will do. But if you want to make reliable forecasts or draw trends, especially if they involve large datasets, then there are much better tools.

This paper will return to consider the answer to the second question – how hard is it to learn? – when it looks at an alternative to spreadsheets for statistical calculations.

Before continuing, however, it is worth noting that people use spreadsheets for tasks other than numerical calculation. For example, spreadsheets are frequently used as if they were databases, to create and manage lists. Again, the principles of scale and complexity apply. Beyond certain limits, a proper database, with built-in rules for structuring data, maintaining data integrity, developing audit trails and so on, is far more suitable.

Two things to remember about spreadsheets

Creating a spreadsheet is as complex and error prone as computer programming.

Spreadsheets are really computer programs

When you design a spreadsheet layout, you are writing a computer program. Spreadsheet programs such as Microsoft® Excel use what is known as a "non-procedural programming language." Although it is also possible to write procedural programs for Excel in Visual Basic, the everyday business of typing formulas into cells is an exercise in non-procedural programming.

Normally, when we think of programming languages, we think of BASIC, C, Java™, FORTRAN and so on. These are all "procedural languages" and each has a coherent methodology that has been developed for programs in such languages. That's because it has become clear over the years that strict adherence to those rules is crucial to getting programs to operate correctly. Even so, it can take an enormous amount of testing and de-bugging to get a complicated program to produce the right numbers.

Non-procedural programming is just as full of decisions, complexity and chances for mistakes as all but the simplest procedural program.

With standard software development methodology, procedural computer programs are double- and triple-checked. By contrast, a spreadsheet, although it may be vitally important to the operation of a company, is usually the work of one person. It is almost never checked or tested in detail, and quite often goes into production with little or no verification. Yet important management decisions – revenue forecasts and plans for future investment, for example – are based on the numbers that it produces

Studies reveal that 90 percent of all spreadsheets contain at least one error.

Spreadsheets are prone to errors

A number of studies have been made concerning the frequency of errors in spreadsheets. Based on these, it seems that 90 percent of all sheets contain at least one error. The studies were carried out by making visual inspections of mission-critical spreadsheets, so it is possible that many other errors were not found. Also, it was found that attempts to correct errors often introduced new ones.

Brown and Gould[1] detected 17 errors in one spreadsheet, of which 15 led to wrong numbers; 11 involved formulas; 2 were mistyped entries; and 1 was a rounding error. In an analysis of 88 spreadsheets, 94 percent were found to contain errors, and a detailed scrutiny of 43 of them revealed that 5.2 percent of the cells contained errors.

Teo and Tan[2] performed an experiment to determine the frequency of errors introduced by people using a spreadsheet. They showed that 30 percent of outlier errors were corrected by those making a modification to an existing spreadsheet. However, offsetting the improvement was an increase in the overall number of errors: 49 percent had new mechanical errors, 30 percent had new logical errors and 65 percent had new omission errors.

A number of other studies have been made, and the results are consistent. All of the studies indicate that 90 percent or more of the spreadsheets studied contained errors. Panko[3] collected evidence from several studies and the lowest error rate he found was 86 percent – and the highest, from another source, was 100 percent.

Types of spreadsheet errors

Spreadsheet errors can be divided into three main types.

The “friendliest” type is what you could call functional errors. These errors are the easiest to find because they simply stop the spreadsheet from working. Instead of giving you wrong numbers, they give error messages, or nothing at all.

Types of spreadsheet errors include functional errors, outlier errors and stealth errors, which range in severity from low to high.

Then there are outlier errors. With these, the spreadsheet appears to work, but the numbers aren't right. Often, such errors are spotted by someone who has an idea of what the results should be and draws attention to the fact that the results don't match expectations.

The worst are what might be called stealth errors. These produce incorrect results, but nobody believes they're incorrect. They pass inspection and are accepted as the truth. Stealth errors occur either because nobody knows what the correct result should be (which is often the case with statistical calculations), or the numbers are only slightly different from expectations and seem reasonable. It can be years, if ever, before these errors are found.

There are a number of stories about how spreadsheet errors have had embarrassing consequences. One concerns Nevada City, California, which in January 2006 discovered that it had a budget deficit of five million dollars. The budget spreadsheet was the same one they had used previously, but when entering the data for the new year, a formula had been inadvertently overwritten. Fortunately, it was an outlier error and quickly noticed by the city councilors. However, it took the finance director an entire day to correct it (and while he was doing so, he found a number of other errors).

Another story, from 2003, is about a university that discovered errors in the averages of some students' grades: The numbers simply didn't make sense. After performing the calculations by hand, examiners corrected the grades and went on to discover that the error in the spreadsheet's equations was caused by cut-and-paste operations that failed to take into account the difference between absolute and relative cell addressing. And although the spreadsheet had been checked by a senior staff member, he had carefully examined just the first row – which happened to be the only one that was correct.

Causes of spreadsheet errors

Spreadsheet users should know what factors cause errors. Unfortunately, there are too many causes to list here, but the major ones are:

- **Mistakes in logic:** These can be something simple, such as calling the wrong function, subtracting instead of adding, or the omission of parenthesis in formula creation. These sorts of errors can also be caused by the implied relationship of the cells in the spreadsheet.
- **Incorrectly copied formulas:** Typing in an equation while reading it from another location often leads to errors, as does cutting and pasting. Copying existing equations to new locations commonly changes the referenced cells, making it important to ensure that the right changes are made in the right way.

Spreadsheet errors occur for many reasons, including logic mistakes, incorrect formulas, accidentally overwritten formulas and misuse of built-in functions.

- **Accidentally overwritten formulas:** A cell that contains an equation looks like a number – all the user sees on first glance is the result. So accidentally inserting a number into a cell already containing a formula will overwrite the equation and turn the contents of the cell into a constant. If other formulas rely on the results from this cell, the error can be compounded significantly.
- **Misuse of built-in functions:** The wrong function can be used – for example, using AVERAGEA, which evaluates text and false entries to zero, instead of AVERAGE, which ignores them. This type of mistake is, unfortunately, very easy mistake to make.
- **Omitted factors:** It is very easy to simply leave something out. This could be an equation, data or both. Errors of this type occur quite often when new data is added to a previously completed spreadsheet. It could be that not all the data is entered, or the some of the new cells aren't included in all the relevant equations.
- **Data input errors:** If you are fortunate, data input errors will lead to outlier errors; but this is not always the case. For example, if 3,5 is entered instead of 3.5, the spreadsheet assumes it is a string rather than a number. The result is that the value zero is used in any formulas referencing that cell. Another mistake would be to enter 3/5, which becomes a date – a huge number in any calculation.

There are a plenty of other possibilities. For example, if you sort a column that contains both figures and equations, you will sort the equations in addition to the figures, which can result in computation errors.

Finally, there is the issue of the reliability of spreadsheet software. Spreadsheet developers continually issues patches and corrections to their software. However, in 2008 Gregg Keizer[4] reported that a patch to correct an error in Excel was found to cause new errors in calculations: In fact, it resulted in more errors than before. (The software had been issued five years previously; but either the issue had not been detected, or, if it had been detected, no attempt had been made to fix it.) It is not just an issue of patching. Several studies show[5] that spreadsheets are simply not very accurate for complicated mathematical procedures, even when they are coded correctly.

Because it is so widely used, even in the teaching of statistics, numerous articles have detailed the errors in statistical procedures in Excel, and more than a few websites highlight its shortcomings with respect to advanced analytics. (See citations at the end of this paper.) To summarize their findings, serious analysis of Excel indicates that it can be downright dangerous to use for statistical analysis. To quote McCullough[5], "Professional statisticians continue to write books with titles like 'Statistics with Excel,' but they now warn students not to bet their jobs on Excel's accuracy."

With spreadsheets, it's easy to make a data entry error, but sometimes extremely difficult to find and fix problems.

Other issues related to using spreadsheets

Entering and checking data

When entering data into a spreadsheet, it's important to be aware of how the information in one cell relates to the cells around it. This means checking that the numbers are within a valid range; that formulas are not over-written by numbers; and that new cells are included in any formulas. It's a case of check, check and check again – and always be ready to hit the Undo button.

It's easy to make a mistake when typing numbers – for example, to enter 95.3 instead of 9.53. If this happens, the spreadsheet will dutifully use the incorrect number, and with luck, the result will be bizarre enough to stand out as an outlier error. Then comes the job of searching the sheet, item by item, to locate the cause (or causes) of the problem.

Accommodating special types of data

There are several types of data, common to many types of research, that require special accommodation.

When recording survey results using spreadsheets, it can be especially difficult to accurately represent missing or categorical data.

One frequently occurring issue is how to handle missing data values. When working with spreadsheets, you have to be careful when dealing with missing values. If you were to assign a zero value to such data, doing so would distort, for example, the average of a range of values. If you enter a string in a cell to indicate a missing value, some equations will ignore the string while others will evaluate it as zero. Because zero is a valid value in some cases, you need another way to indicate a missing value. When doing this, however, not only must you take care to use the missing value designation consistently when entering data, you must also document your approach carefully so that later modifications to the spreadsheet don't invalidate the data conventions you have used. And note that none of the above approaches provides an approved method for reliably imputing missing values.

Another special situation arises when dealing with categorical data (often encountered in survey results). For example, suppose the four values 1, 2, 3 and 4 are assigned to represent the answers "Yes," "No," "Don't know" and "Refuse to answer" in a survey question. If you use a spreadsheet to store this type of data, you have to make a special effort to document the values and what they mean, to ensure that the data is entered correctly (the correct value assigned to each answer) and that it is then processed meaningfully. Otherwise, the meaning will be lost as soon as the person who developed the spreadsheet moves on.

Projecting the future

Typically, spreadsheets are used to extract information and relationships about past events. Increasingly, however, organizations want to know what is likely to happen in the future. Recent versions of Excel have dedicated functions, for example, FORECAST, TREND and GROWTH, for predicting new values based on existing data and a range of plug-in programs are also available. What is in question,

however, is the reliability and accuracy of these capabilities, which, in any case, provide none of the tests that serious mathematicians would use to check the validity of the results.

Though some spreadsheet applications contain functions for predicting future trends and outcomes, these methods are often unreliable and inaccurate.

Data management

With their cell-level focus, spreadsheets present a number of challenges in managing data. A conceptually simple change – such as modifying a start time, adding new members or changing a formula – can require dozens, even hundreds, of other changes.

Even one simple modification can require cell/row/column insertions or deletions, editing or copying formulas across a range of cells, or re-configuring the entire spreadsheet. These operations are not only time consuming, they can actually lead to more errors.

Almost invariably, new data has to be added to a finished spreadsheet. But how should the new figures be accommodated? One way is to set the spreadsheet to extend equations to include all new data: the problem with this is that some equations could inadvertently be extended to include data that shouldn't be in them. On the other hand, if the sheet is set not to extend automatically, some data that should be included is likely to be left out. Either way, it is unlikely that the spreadsheet will produce correct results unless the changes are checked carefully.

An alternative to spreadsheets

So far, this paper has provided an overview of situations in which spreadsheets may prove inadequate, or at least cumbersome, for statistical analysis. This is not to say that they are without value. If the task is to carry out simple tests on a small number of variables, then a spreadsheet is as good a tool as any.

That said, a spreadsheet program is, as stated previously, general purpose software. With or without plug-ins, the range of analytical tools is limited and the spreadsheet program's algorithms are not as rigorously designed or tested as those in software programs specifically designed for statistical analysis.

IBM SPSS Statistics offers organizations the ability to do robust, in-depth statistical information analysis without doing any programming.

Just as a carpenter might use a handsaw to cut up a dozen lengths of wood but would turn to specialized tools for cabinet making and power tools to handle enough lumber for a building, anyone wishing to do robust and in-depth analysis should use a tool especially built for the job. One of these is IBM® SPSS® Statistics* from SPSS, an IBM Company.

IBM SPSS Statistics has been continuously developed and tested since 1968. Over that period, many forms of statistical analysis have been embedded in the software and the algorithms that execute the equations have been tested by both developers and users in academia, in

**IBM SPSS Statistics was formerly called PASW® Statistics.*

laboratories and in virtually every type of business. As a result, users can have confidence that the software has been thoroughly tested and its results found to be reliable.

Without doing any programming, users can run a very broad range of statistical analyses. In addition, as users' understanding progresses, they are immediately able to apply more advanced methods because they are already there, in the software.

Naturally, IBM SPSS Statistics is optimized to handle statistical calculations in a way that a spreadsheet could never be. In fact, the software is optimized for statistical work at every point, from data entry through to the creation of reports for decision makers.

The built-in data validation and error-checking mechanisms of IBM SPSS Statistics help ensure that data entered is valid and correct.

Data entry, the IBM SPSS Statistics way

With IBM SPSS Statistics, the data entry process starts with definitions of the data types that are going to be used. These are quite detailed. For example, every data type has both a long and a short name. (The name that is the best fit is the one used to annotate tables and graphs.) In addition, the type of data that can be entered – numbers or text, to give a simple example – can be specified. At this point, the first level of error checking takes place. The data must fit the characteristics of the defined type, or it won't be accepted. Neither the data type nor any other characteristics of the layout can be modified accidentally. Nor can you alter relationships among the data. Data entry is simply data entry: it is in no way intermixed with programming.**

The data validation and error checking mechanisms supplied with IBM SPSS Statistics are really quite comprehensive. Automatic procedures locate values that appear to be out of line – which takes care of locating most typos. If, however, the value happened to be within range but somehow abnormal when compared to the other numbers entered, IBM SPSS Statistics would spot it and ask about it.

Preparing data for analysis: the IBM SPSS Statistics approach

As noted above, it often happens that the data available for analysis is incomplete. In a survey, for example, some people may miss or choose not to answer a question. As was pointed out earlier, handling incomplete data in a spreadsheet poses numerous difficulties. With the IBM SPSS Statistics, researchers can examine the available data and calculate values for missing items (a process called "imputation"). They can examine data using one of six diagnostic reports to uncover missing data patterns.

*** Note: SPSS, an IBM Company, recommends using a special-purpose product such as IBM® SPSS® Data Collection Data Entry or another product in the company's survey research software product line to enter data, as these are designed to check for errors as data is entered. IBM SPSS Data Collection Data Entry was formerly called PASW® Data Collection Data Entry.*

Or they can estimate summary statistics and impute missing values, using an automated procedure that chooses the most suitable imputation method based on the characteristics of the data. Then the analysis can be performed just as if all the data were present – which, in a very real and mathematically valid sense, it is.

Other data steps in preparing data for analysis include looking at the distribution of data, checking for outliers and organizing or "binning" data so that the algorithms you plan to use – such as Naïve Bayes or logit models – operate efficiently. IBM SPSS Statistics performs these data preparation steps – something no spreadsheet program is designed to do.

Statistical analysis with IBM SPSS Statistics

Data analyzed with IBM SPSS Statistics may be output to various formats including an impressive array of plots and graphs.

When IBM SPSS Statistics switches to analysis mode and performs the actions necessary to produce an output, the data is not modified: it is used only as input to the process, and the output – available in various formats including an impressive array of plots and graphs – is displayed in a separate window.

In addition, when any type of analysis is performed, the software automatically writes a program in the form of syntax that can be saved and run time after time on different datasets without the need to change it. (It can be changed if required, though.)

IBM SPSS Statistics also has the advantage of allowing advanced users to implement new procedures and functionality through its Programmability Extension. This advanced feature allows users that are comfortable with the R statistical programming language, or with Python®, to embed new algorithms or functions directly into the product. They can even create a native GUI for the new feature they've created to give access to it to non-programmers, who can then perform the analyses on their own, rapidly and efficiently.

Looking into the future with IBM SPSS Statistics

Though spreadsheets can be used to make some kinds of projections, you need a tool such as IBM SPSS Statistics to factor in complex variables.

Spreadsheets are frequently used to make projections – to estimate future events based on historical data. For example, a common business application would be to forecast the next two quarters' revenues based on the previous year's results. Although it is possible to carry out such a calculation using a spreadsheet, accounting for factors such as seasonality in a business, or developing scenarios based on multiple other variables is only feasible with mathematically robust software such as IBM SPSS Statistics.

The learning curve with IBM SPSS Statistics

At the beginning of this paper, we posed a question that often rears its head when people are considering using new software: How hard is it to learn?

In the case of IBM SPSS Statistics, the answer is: "Not hard at all." Like a spreadsheet, it has a WYSIWIG interface, so everything is laid out clearly to view and its features are accessed through familiar Menu and Toolbar layouts. The program's statistical functions are logically grouped: When one is selected, the relevant options appear in a pop-up window and the calculation is performed by choosing the required options and clicking the "OK" or "Run" button.

IBM SPSS Statistics is an easy-to-learn tool that helps your organization move beyond spreadsheets to mathematically robust analysis of complex data.

In addition, IBM SPSS Statistics comes with very comprehensive tutorial, extremely detailed Help files and case studies that detail examples of the use of statistical analysis in business and research situations. Together, these can take a user from statistical novice to competent analyst fairly quickly. The company, of course, offers a number of training options, including on-demand web-based training. Also, because of its long history of use among analysts in all types of settings, additional learning resources are available from third parties, including online discussion boards with tips from other users, instructional books and videos, textbooks and workbooks.

Conclusion

In writing this paper, the author discovered several things worth bringing up at this point: First, spreadsheets are used far more widely than is generally realized – often without seeking out other solutions. Second, the error rate in spreadsheet usage is astounding, particularly when compared to the acceptable error rate in other forms of computing. Third, spreadsheets are called upon to address a wide variety of problem types – some of which may not be at all suited to the programs' capabilities.

Your dataset is unique and so is the way you might use a spreadsheet.

For you to understand whether a spreadsheet is sufficient for your needs or if you might benefit from a specialized tool such as IBM SPSS Statistics, it's best to see for yourself how each program works with your data, carrying out the analytical tasks you normally require.

It's easy to test drive IBM SPSS Statistics; you can contact the company or download a free evaluation copy of the software at www.spss.com/statistics. If your data is already in a spreadsheet, IBM SPSS Statistics can easily import it. And once your data is in, you can evaluate the types of analysis available and see if there are advantages, in your case or in some situations, to using a tool designed for statistical analysis, in place of a general-purpose spreadsheet program.

About IBM Business Analytics

IBM Business Analytics software delivers complete, consistent and accurate information that decision-makers trust to improve business performance. A comprehensive portfolio of business intelligence, advanced analytics, financial performance and strategy management and analytic applications gives you clear, immediate and actionable insights into current performance and the ability to predict future outcomes.

Combined with rich industry solutions, proven practices and professional services, organizations of every size can drive the highest IT productivity and deliver better results.

For more information

For further information or to reach a representative:
www.ibm.com/software/analytics/spss

Request a call

To request a call or to ask a question, go to **www.ibm.com/software/analytics/spss/contactus**. An IBM representative will respond to your inquiry.

Notes

1. P. Brown, J. Gould, "An experimental study of people creating spreadsheets," *ACM Transactions on Office Information Systems*, (1987) Vol. 5, 258-272.
2. T. Teo, M. Tan. "Quantitative and qualitative errors in spreadsheet development," *Proceedings of the 30th Hawaii International Conference on Systems Sciences* (1997) 149-155.
3. Ray Panko, Professor, University of Hawaii. www.panko.cba.hawaii.edu/ssr.
4. Gregg Keizer, "Microsoft fixes Excel math mistake," *Computerworld* (March 2008).
5. Bruce D. McCullough, "The Unreliability of Excel's Statistical Procedures," *Foresight*, (February 2006) Vol. 3, 44-45.

Other sources

www.daheiser.info/excel/frontpage.html

www.practicalstats.com/xlsstats/excelstats.html

McCullough, B.D. and Wilson, B. "On the accuracy of statistical procedures in Microsoft Excel 2003." *Computational Statistics and Data Analysis.* (2005) Vol. 49, 1244-1252.

– "Teaching statistics with Excel 2007 and other spreadsheets." *Computational Statistics and Data Analysis*, (June 2008) Vol. 52, issue 10

IBM Corporation
Route 100
Somers, NY 10589

Produced in the United States of America
May 2010

Please Recycle

Business Analytics software

IMW14297USEN-02