Practical Stats Newsletter for October 2009

Subscribe and Unsubscribe: http://practicalstats.com/news
All of our past newsletters:
http://www.practicalstats.com/news/news/bydate.html

In this newsletter:
1. Upcoming Courses
2. Excel and Statistics
3. Give us your ideas for the new edition of the nondetects textbook

1. Upcoming Courses
Classes now scheduled in 2010:

Time Series and Forecasting (for frequently-collected data)
Kilauea Military Camp, Hawaii Volcanoes Natn. Park (the Big Island)
Hawaii. Feb. 3-4, 2010.
This class has a very limited enrollment, so I'm announcing it here before I go 'public'. Sign up sooner rather than later to insure that you get a seat.

Time Series and Forecasting (for frequently-collected data)
Golden, Colorado Feb. 16-17, 2010.
Nondetects And Data Analysis
Golden, Colorado Feb 18-19, 2010.

You can always find a complete course listing at
http://www.practicalstats.com/new_classes/classes.html

2. Excel and Statistics

Perhaps the most popular page on our website over the years has been our "Statistics With Excel?" page. It discusses the problems with using Excel for statistics, given that Excel isn't a statistics software package. Two recent items on the topic are noteworthy, and our web page has been updated in response.

A) B. McCullough has co-authored a series of papers over the years evaluating the accuracy (and lack of it) for MS Excel's statistical routines. In 2008 he collected several papers into a special issue of the journal *Computational Statistics and Data Analysis* (vol. 52, issue 10) evaluating the routines in Excel 2007. It appears that not much has changed in the Excel world. Problems exist in several statistical functions and routines that are likely used by any scientist who uses Excel. Tests may return inaccurate p-values. Perhaps the most egregious error is in Excel's normal probability plot produced for regression residuals. This plot incorrectly compares data to a uniform distribution, not a normal distribution, so the evaluations of whether data follow a normal distribution are completely bogus. His conclusion: "…it is perhaps worth comparing quality

assurance in Microsoft's Excel to quality assurance in its game division ... It is difficult not to think that if Microsoft tested business software the way it tested game software, then the statistical functions in Excel would be as accurate as those found in any other major software package. If that were the case, then none of the articles in this special section would have been written."

Some highlights of the papers include

McCullough and Heiser (2008), pages 4570-4578:
"... it is not safe to assume that Microsoft Excel's statistical procedures give the correct answer. Persons who wish to conduct statistical analyses should use some other package." Specifically,
1. "a major flaw in Excel Solver" which is used to solve nonlinear equations. Similar flaws in other nonlinear functions such as LOGEST.
2. "the Excel random number generator does not fulfill the basic requirements ... for scientific purposes."
3. The normal probability plot of residuals in regression analysis is totally wrong. "Excel computes the plot for the wrong variable, the dependent variable instead of the residuals, and has managed to confuse the uniform distribution for the normal distribution."
4. Inaccurate t-test results in the presence of missing values. (I was unable to verify this, but that is what they state)
5. Inaccurate p-values from a t-test.
6. The trendline function computes incorrect regression equations and can produce a negative r-squared.
These problems are "by no means exhaustive".

Yalta (2008), pages 4579-4586:
The accuracy of the binomial, Poisson, inverse standard normal, inverse beta, inverse student's t and inverse F distributions is very poor. The inverse distributions are used to produce p-values for hypothesis tests. In contrast, the equivalent results in the free OpenOffice's Calc spreadsheet and the open-source Gnumeric spreadsheet are superior. A table of numerical errors reported for Excel 97 is carried through for later versions. Almost all are listed as "Not fixed" or "Poor fix" in Excel 2000, 2003 and now in 2007. The summary: "researchers should continue to avoid using the statistical functions in Excel 2007 for any scientific purpose."

Su (2008), pages 4594-4601:
Graphics in Excel have defaults that lead to chartjunk – "redundant symbols, fill-ins and other extraneous graphical elements" that don't tell readers anything about the data themselves. It is possible to make good graphs with Excel, graphs that adhere to current principles of good graphics, but these require the user to invoke options. So that requires users to be familiar with principles of good graphics and override the defaults.

More detail is found on our Statistics with Excel? webpage,
http://www.practicalstats.com/xlsstats/excelstats.html

There you will find three possible solutions to the problems:
1) Purchase Excel statistical add-ins

2) Use free, open-source alternatives that do not have the issues found in Excel. There are alternative spreadsheets out there, for free.
3) Purchase commercial software, or use the free R statistical software system.

B) Richard Heiberger and Erich Neuwirth have published a book titled "R Through Excel: A Spreadsheet Interface for Statistics, Data Analysis, and Graphics", part of the UseR series. Their book presents their Excel add-in called RExcel, that allows you bypass Excel statistics functions and instead call R routines from within Excel. The results, including graphs, are returned to your Excel worksheet. I've seen their presentation on the add-in, and it is impressive. The primary reason people have continued to use Excel for statistics is that it was already on their hard drive, and so "free". Given that R is truly free, linking the two provides world-class statistics routines that can be accessed from within the familiar Excel environment, for free. Between this add-in and the other free, open-source software packages mentioned above, there is no reason to live with Excel's substandard subset of routines anymore. RExcel can be downloaded for free using the link on our page http://www.practicalstats.com/xlsstats/excelstats.html
Also see our Summer 2006 newsletter on R at
http://www.practicalstats.com/news/news/bydate.html). A textbook 'users manual' for RExcel is published by Springer, and available from the usual sources of technical books.

3. Give us your ideas for the new edition of the nondetects textbook

I'm beginning to write the second edition of Nondetects And Data Analysis, (NADA), the textbook that discusses methods for handling nondetect data. I'd like to get your ideas on what should be added, changed, or deleted from the current book. Here's my plans:
1. add a new chapter on multivariate methods with nondetects. How to do principal components, cluster analysis, etc. when some data are nondetects.
2. add a new chapter on R. More detail on the NADA for R package and how to do all these methods, without the annoyance of 'flipping' data now required by commercial software, within the free, internationally-standard R statistical package.
3. add more detail on both parametric and nonparametric methods when coding values truly below the detection limit differently from values lying between the detection and quantitation limits.
4. add a new introduction along the lines of my paper "Fabricating Data" doi:10.1016/j.chemosphere.2006.04.051
5. add a section on how to sum a series of values, some of which are nondetects. The primary application is to things like dioxin congeners and their TECs, TEQs, etc. This information is in my paper "Summing Nondetects", now accepted for publication by the journal Integrated Environmental Assessment and Management (SETAC).

If you would like to give me your ideas on what to change/add for the next edition of the NADA book, go to our blog page
http://www.practicalstats.com/blog/
and add you comments. Please keep it clean. No ads.

Or use the “Contact Us” tab to send me an email. I welcome any technical comments you have.

'Til next time,

Practical Stats (Dennis Helsel)
-- Make sense of your data