

Practical Stats Newsletter for August 2008

Subscribe and Unsubscribe: <http://practicalstats.com/news>

Archive of past newsletters: <http://practicalstats.com/news>

In this newsletter:

1. Upcoming Courses
2. Review of Lower-Cost Statistical Software, Part 1
3. Visit us at SETAC 08

1. Upcoming Courses

There are several courses coming up in the next few months. You can always find a complete listing at http://www.practicalstats.com/new_classes/classes.html.

Nondetects And Data Analysis, the course that illustrates methods for correctly handling data with nondetects, will be held at the Mainsail Suites and Conference Center in Tampa, FL on Nov. 11-12, 2008. Registration is \$795 online at http://www.practicalstats.com/new_classes/classes.html. The Conference Center is near the Tampa International Airport, so no car rental (and its associated gas prices) are necessary. Special rates on hotel suites are also available - see the above registration link.

Immediately following on Nov. 13-14 at the same location is *Untangling Multivariate Relationships*. This course covers the multivariate methods of primary interest to environmental science, focusing on what each method is designed to do, when to use them, and when not to. More detail on course content is on our website. Registration is \$895 through the same link, above. The following week in Tampa is the SETAC (Society for Environmental Toxicology and Chemistry) annual meeting.

The one-day seminar *Introduction to Practical Statistics* will be held on Sept. 24, 2008, just before the California Groundwater Resources Association annual meeting in Costa Mesa CA. It will introduce some of the concepts in our *Applied Environmental Statistics* course, looking at them from the perspective of groundwater quality. For more information, see <http://www.grac.org/stats.asp>. If you know a ground water scientist who has been alienated by statistics, this workshop will begin the reconciliation.

2. Review of Lower-Cost Statistical Software, Part 1

This month we review five lower-cost options for performing statistical methods. These five and other programs are linked in various ways to Microsoft Excel. As stated in our review of more traditional statistics software {see the Nov 07 Practical Stats newsletter at <http://www.practicalstats.com/news>}, the cost of commercial stat software is high, typically \$1500 and sometimes more. For scientists without access to a corporate discount, this is quite high. Yet environmental scientists have sophisticated needs. How many necessary procedures, such as regression diagnostics for building a good multiple

regression equation, can be found in lower-cost software? Next month we continue by reviewing additional lower-cost software in a special edition of our newsletter (usually we send out newsletters only once per two months).

The five programs reviewed this month range in price from \$50 to \$445. The five are

Fast Statistics	v.2	www.fatesoft.com/excel	\$50
Statisti-XL	1.8	www.statistiXL.com	\$75
WinStat		www.winstat.com	\$99
Analyze-It ME		www.analyze-it.com	\$185
xlStat	2008.5	www.xlstat.com	\$445

We tested the Windows version of each package running MS Vista with Excel 2003. A Macintosh version of xlStat is also available. A pdf file with our complete breakdown of the feature set of each package is available at

<http://www.practicalstats.com/news>

All five packages perform several basic statistical procedures with a menu-driven system. Some run as macros, adding a toolbar or menu within Excel. Others are standalone applications that read Excel files. All estimate percentiles, means and other summary statistics. All perform t-tests (paired and 2-sample), the Mann-Whitney, Kruskal-Wallis, and signed-rank tests. Note that in order to get the Mann-Whitney test in Fast Statistics you must perform the Kruskal-Wallis test procedure on two groups of data. All perform ANOVA and estimate regression slopes and intercepts. All compute Pearson's r correlation coefficient and draw scatterplots. All compute contingency table (chi-square) tests on a table of counts.

From there the feature sets of the packages diverge, with the more expensive packages generally containing more features. Fast Statistics does not perform many functions necessary for analysis of environmental data. It cannot plot boxplots by groups on one plot, one of the most helpful procedures for comparing among groups of data. It cannot compute Kendall's tau correlation coefficient, the basis of several tests for trend. It cannot test for differences in variance (lack of precision) by groups. It cannot perform multiple comparison tests as a follow up to ANOVA or Kruskal-Wallis. It has no regression diagnostics such as Mallows' Cp, adjusted r-square or VIFs. It cannot plot partial plots. It does however compute regression residuals, which can be copied and pasted back into the Excel worksheet and then plotted to produce residuals plots (albeit with a great deal of work). However, it only includes the crude chi-square test for normality, so judging the distributional assumption of regression residuals, or of any original set of data, is not really possible. In sum, its feature set is inadequate for even basic analysis of environmental data.

StatistiXL and WinStat add a few more features for a small increase in price. Both perform Tukey's multiple comparisons and Spearman's rho correlation. For some reason both perform the multivariate methods of discriminant function and factor analysis. StatistiXL displays residual versus predicted plots for regression, and is the only one of the five to perform partial plots. However, StatistiXL performs no tests for normality,

while WinStat adds only the KS test, not a powerful test for judging normality of continuous data. WinStat does compute Kendall's tau correlation, but provides no way to test one-sided alternatives in its hypothesis tests. In short, these two packages also come up short for scientific applications. StatistiXL is relatively easy to use, and if it added Kendall's tau and some regression diagnostics such as VIFs and adjusted r-squared, it could be a useable low-end scientific statistics package. But not at present.

Analyze-It and xlStat add a considerable number of features for their additional (though still comparatively reasonable) costs. Both programs add better normality tests (Shapiro-Wilk and Anderson-Darling) and easily allow one-sided alternatives. Both compute Kendall's tau correlation and include residuals plots for multiple regression. On top of this xlStat performs Levene's test for differing variances, a more modern and accurate test than the traditional Bartlett's test. Only xlStat computes VIF statistics for regression, and Dunn's nonparametric multiple regression procedure to follow up the Kruskal-Wallis test. xlStat performs logistic regression and principal components analysis. It is the only package of the five that computes a lowess smooth and saves residuals from it, a handy tool in trend analysis.

None of the five packages perform some helpful functions. None compute prediction or tolerance intervals for a column of data. None delve into any variation of bootstrapping or permutation tests. None perform any power or sample size analysis. None of them compute the Sen slope, the trend slope for the Mann-Kendall test for trend. Only StatistiXL computes partial plots, which are incredibly important in building regression models. None of them perform equivalence tests.

In short, none of these five packages can be entirely recommended as a replacement for a full-fledged statistics software package when analyzing environmental data. The most expensive, xlStat, could be considered sufficient and useful for its price if it added the capability for partial regression plots. Next month we'll look at some other alternatives, in our "Man vs. Stats" attempt to survive in the desert of environmental statistics.

3. Visit us at SETAC 08

Practical Stats will be exhibiting at the SETAC (Society for Environmental Toxicology and Chemistry) national conference in Tampa in November. Stop by, say hello, and pick up our free CD of software and more.

'Til next time,

Practical Stats

-- Make sense of your data