

Resources for Learning about Statistics & Data Analysis

A List of Resources for California State and Regional Water Board Staff¹

January 12, 2009

A. Water Board Training Academy Courses

In the past, the Training Academy has sponsored a five-day Applied Environmental Statistics course (2003, 2005) and a two-day course on Statistical Methods for Data Below Detection Limits (2003). These courses are taught by Dr. Dennis Helsel, formerly with the USGS, along with hands-on training with MINITAB statistical software.

Scheduling of this type of training is based on a communicated need by Water Board staff to the Training Academy. No courses are planned for 2009.

B. Intranet

The Division of Water Quality maintains an extensive Technical Library on the intranet site. The “Data Analysis” portion of this e-Library contains entire statistical textbooks, journal articles, pamphlets, PowerPoint presentations, and software links. The intranet site is:
http://waternet/dwq/pubs/html/data_analysis.html

C. Internet sites

Many informative internet sites are available for learning about statistics and data analysis methods. Some useful sites are:

<http://www.practicalstats.com>
<http://www.quantdec.com/envstats/index.htm>
<http://wiki.stat.ucla.edu/socr/index.php/EBook>
<http://www.statisticalengineering.com/index.html>
<http://linkage.rockefeller.edu/wli/glossary/stat.html#beta>
<http://davidmlane.com/hyperstat/index.html>
<http://www.itl.nist.gov/div898/handbook/index.htm>
<http://www.itl.nist.gov/div898/software/dataplot/>
http://www.scisoftware.com/products/chemstat_details/chemstat_details.html
<http://www.claviusweb.net/statistics.shtml>
<http://cran.r-project.org/>

¹ Compiled by Steve Saiz, ssaiz@waterboards.ca.gov, Central Coast Regional Water Board,

D. On-line Video Instruction in Statistics

Against All Odds Inside Statistics

After registering, you can watch *Against All Odds – Inside Statistics*, a video instructional series on statistics for college and high school level students at <http://www.learner.org/resources/series65.html>. With an emphasis on “doing” statistics, this series goes on location to help uncover statistical solutions to the puzzles of everyday life. Learn how data collection and manipulation — paired with intelligent judgement and common sense — can lead to more informed decision-making. This series, dating from 1989, consists of 26 half-hour video programs having the following titles:

1. What Is Statistics?	14. Samples and Surveys
2. Picturing Distributions	15. What Is Probability?
3. Describing Distributions	16. Random Variables
4. Normal Distributions	17. Binomial Distributions
5. Normal Calculations	18. The Sample Mean and Control Charts
6. Time Series	19. Confidence Intervals
7. Models for Growth	20. Significance Tests
8. Describing Relationships	21. Inference for One Mean
9. Correlation	22. Comparing Two Means
10. Multidimensional Data Analysis	23. Inference for Proportions
11. The Question of Causation	24. Inference for Two-Way Tables
12. Experimental Design	25. Inference for Relationships
13. Blocking and Sampling	26. Case Study

E. Statistical Software

Simple analyses and data summaries can be made in Microsoft Excel; be sure to install the free *Analysis Tool Pack*. (In Excel, click on Tools/Add-Ins.) A free Excel worksheet for computing summary statistics of data containing non-detected observations (i.e., censored data) using the Kaplan-Meier method is available from

<http://www.practicalstats.com/nada/nadafiles/downloads.html>.

Dedicated statistical software such as MINITAB, SPSS, SYSTAT, and SAS allow more complex analyses to be easily performed without programming. These higher end software packages also provide good graphics for presenting data. However, dedicated statistical software is expensive, usually starting at \$1,500.

Lower-priced statistical software is available. The Practical Stats website

<http://www.practicalstats.com> reviewed nine statistics programs ranging between \$50 and \$599. Some of these are Excel add-ons. I have reproduced the review in Appendix 1.

Free statistical software is also available. Many free macros are written for MINITAB. ProUCL 4.0 was written by USEPA for the statistical analysis of environmental data sets with and without

nondetect observations <http://www.epa.gov/esd/tsc/software.htm>. — You can access or download the excellent, informative ProUCL Technical Guide and Users' Manual from the DATA ANALYSIS topic listing in the Water Board's Technical e-Library, mentioned above.

Another free statistical software is RPCalc 2.0, which was written by SWRCB-DWQ for estimating summary statistics and upper tolerance bounds of environmental data with and without nondetects according to the California Ocean Plan reasonable potential procedure http://www.waterboards.ca.gov/water_issues/programs/ocean/docs/oplans//rpscalc.zip.

Two main water quality data analysis packages are:

- DUMPStat [<http://www.dsi-software.com/>]. DUMPStat is a computer program for the statistical analysis of groundwater monitoring data using methods described in [Statistical Methods for Groundwater Monitoring](#) by Dr. Robert D. Gibbons; and
- Sanitas [<http://www.sanitastech.com/sanitas/sanitas.html>]. Each year there is a two-day Sanitas class in California and, after attending, regulators get a copy of the software, plus any necessary consulting (e.g., for problems importing the data set into Sanitas format), without charge.

F. Statistical Consulting

The State Water Board maintains an on-going contract with the UC Davis Statistical Laboratory (Stat Lab) for consulting services by State and Regional Water Board staff. The Stat Lab serves as a professional resource to external agencies and individuals whose work includes applied statistical modeling and inference. Inquiries at the early stages of statistically based investigations are particularly welcome. This includes questions on the design of experiments or sample surveys and the assembling and management of databases.

Most Water Board staff will consult with Dr. Neil Willits, Ph.D., Senior Statistician. Dr. Willits is well versed in a wide range of statistical methodologies and has worked on many long-term and large-scale projects with external agencies, including the California Water Resources Board, the State Departments of Toxic Substance Control, and California Department of Fish and Game.

Be sure to contact William Ray, the SWRCB contract manager, prior to consulting with the Stats Lab in order to obtain the correct billing code.

William (Bill) Ray Quality Assurance Office Manager California SWRCB Office of Information Management and Analysis (OIMA) 1001 I St., 15th floor PO Box 100 Sacramento, CA 95812-0100 (916) 341-5583 bray@waterboards.ca.gov	UC Davis Statistical Laboratory One Shields Avenue 4118 Mathematical Sciences Building Davis, California 95616 Ph: 530.752.2361 Fax: 530.752.7099 http://www.stat.ucdavis.edu/stats-lab/services
---	---

Appendix 1. Review of Lower-Priced Statistical Software²

Part 1 (from August 2008 Newsletter)

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/aes/aesfiles/DownloadsAES.html>

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

² Adapted by Steve Saiz from "Practical Stats Newsletter" for August and September 2008, available at <http://practicalstats.com/news>.

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.

Part 2 (from September 2008 Newsletter)

This month we review four lower-cost options for performing statistical methods, alongside the five reviewed in our August newsletter. An Excel spreadsheet evaluating the feature set of all nine software packages is available on our newsletter site, [<http://www.practicalstats.com/news>]. Environmental scientists have sophisticated statistical needs. Some of the features we evaluated each package for, and which for some came up lacking, include residuals and partial plots for multiple regression, regression diagnostics such as Mallow's Cp and VIFs, better tests for normality such as Shapiro-Wilk, Anderson-Darling or PPCC (not the much older Kolmogorov-Smirnoff or chi-square tests), and both parametric and nonparametric multiple comparison tests. These are all elements of what we teach in our Applied Environmental Statistics course, and should be tools of the trade when examining environmental data.

The four programs reviewed this month range in price from \$200 to \$599. The four are:

StatPlus	www.analystsoft.com/en/products/statplus/	\$200
WinksProfessional	www.texasoft.com	\$229
StatTools	www.palisade.com/stattools/	\$595
NCSS	www.ncss.com	\$599

We tested the Windows version of each package running MS Vista, except for Winks, which we ran on Windows XP. A Macintosh version of StatPlus is also available. If you haven't seen last month's review of the first five packages, it is available on our newsletter page, www.practicalstats.com/news.

All four packages estimate percentiles, means and other summary statistics. All perform t-tests (paired and 2-sample), the Mann-Whitney, Kruskal-Wallis, and signed-rank tests. 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, though we could not get Stat Plus to return correct results for contingency tables.

From there the feature sets of the packages diverge. NCSS provides the most extensive feature set of all of the nine packages we reviewed. It performs all of the features we were looking for in an environmental statistics package, except for Lowess smoothing and Kendall's tau correlation with the associated Theil-Sen line. Even there, NCSS performs an alternative robust line method, and so provides the functionality of Theil-Sen. Its capabilities best match the features we were looking for among all the nine software programs tested here. It includes modern regression diagnostic methods, the newer tests for normality, and both parametric and nonparametric multiple comparisons. For an individual scientist without a corporate license for one of the 'major' statistics packages costing \$1500 and up (see our November 07 newsletter on the cost of statistical software), NCSS would provide a complete suite of methods for much lower cost.

Stat Tools feature set is more consistent with software in the \$200-\$300 range, such as the two tested this month, Winks and StatPlus. Each of these three include and exclude methods in

slightly different areas, but tend to not include some of the modern regression diagnostics that help scientists build good multiple regression models. Of the three, only StatPlus computes the Shapiro-Wilks test or one of the other better methods for judging how closely data fit a normal distribution. Only StatPlus computes Kendall's tau correlation coefficient. None of the three compute a nonparametric multiple comparison test.

Check the full feature set evaluation on our newsletter site, to see which software package includes the features you require. In addition, a short summary evaluation of the nine packages is below, based on their ability to perform the procedures we find most necessary for the analysis of environmental data. The maximum rating is 6★s. Six stars seemed appropriate since the maximum cost was \$600. One could look in your price range and find a package that had at least one star per \$100. In our judgment, to use one of these software packages as your only statistics package for environmental applications, you would need a package rated at either 5 or 6 stars.

NCSS	\$599	★★★★★★
Stat Tools	\$595	★★
xlStat	\$445	★★★★★
Winks Pro	\$229	★★
Stat Plus	\$200	★★
Analyze It	\$195	★★★
WinStat	\$99	★
StatistiXL	\$75	★★
Fast Stats	\$50	★