Practical Stats Newsletter   for Spring, 2004

In this newsletter:
1. Upcoming Courses
2. Computing the mean (or median) of data with nondetects
3. Nondetects And Data Analysis textbook completed!
4. Next quarter's newsletter

1. Upcoming Courses

Less Than Obvious, our 2-day course on the analysis of data with nondetects, will be taught again on August 18-19, 2004 on the campus of the Colorado School of Mines in Golden, Colorado. Registration information is on the PracticalStats web site. Course content follows the new NADA textbook (see below).

Our one-week survey course of practical statistics for natural resources, Applied Environmental Statistics, will be taught in Golden, on June 7-11, 2004. This is how to "make sense of your data". An outline of course content is on the PracticalStats web site. It covers topics from plotting data and hypothesis tests to regression, trend analysis, and comparing data to standards.

We're booked until fall, but could teach a course in Oct-Dec at other locations - if you know of at least 10 people ready to take either course we'll arrange to come to your area and fill the remaining slots through advertising. Or we'd be glad to teach it just for your group. Contact us at ask[at]practicalstats.com.

2. Computing the mean (or median) of data with nondetects

More articles have been written comparing and recommending methods to compute summary statistics than for any other type of analysis of censored environmental data. As early as 1967, Miesch (1967) recommended use of maximum likelihood estimation (MLE) for computing estimates of mean abundances (mass) of metals with censored measurements in rock samples. Discussions and recommendations continue to this day, with a myriad of methods used in areas such as water quality, air quality, soil contamination, and others. Each year seems to bring a new guidance document, with conflicting recommendations from the previous ones.

In Chapter 6 of the new book Nondetects And Data Analysis I recommend three methods for computing summary statistics of censored environmental data, with the choice a function of the number of observations and the percent of the dataset which is censored:

Recommended methods for estimation of summary statistics

	Amount of available data	
Percent Censored	< 50 observations	> 50 observations
< 50% nondetects	Kaplan-Meier	Kaplan-Meier
50 - 80% nondetects	robust MLE or ROS	Maximum Likelihood
> 80% nondetects	report only % above a meaningful threshold	may report high sample percentiles (90th,95th)

For datasets with less than 50% censoring I recommend the Kaplan-Meier estimator. This method is unfamiliar to many environmental scientists. What is it, and why use it?

Kaplan-Meier is the standard method for handing censored data in the medical sciences, where censored values occur as 'greater thans'. It is easy to convert the method for use with the 'less-thans' of environmental measures ˇ see Chapter 6 in NADA. Kaplan-Meier is the nonparametric maximum-likelihood method for computing summary statistics. In other words, it is the most efficient method available for censored data that does not assume that the data follow any specific distribution. Kaplan-Meier estimates the cumulative distribution function (cdf or edf) of a dataset, incorporating censored values at one or more reporting limits. >From the cdf an estimate of the moment statistics mean and standard deviation can be derived (though they are of lesser importance in medical statistics, where the use of percentiles dominates). If there were no censoring, K-M produces the familiar sample estimates for mean and percentiles At close to 50% censoring, Kaplan-Meier does not produce an estimate for the median, and a model-based method that makes some type of distributional assumption, at least for censored values, must then be employed. Kaplan-Meier is the standard procedure in a large discipline that has successfully dealt with censored data for decades. We in environmental sciences ought to pay attention to that as we decide what methods might become standard for our use.

As censoring increases, so-called 'robust' methods employ a distributional assumption for the censored portion of a dataset, while traditional maximum likelihood methods do so for the entire distribution. Authors of articles comparing methods for computing summary statistics with censored data have usually found that maximum likelihood methods do not work well for small datasets. With fewer than about 50 observations, the method can be fooled by one or two outliers, producing inaccurate summary statistics. In addition, if the summary stats are computed following transformation to another scale, say with logarithms, and the statistics retransformed back into original units, a transformation bias is introduced that is difficult to overcome with small data sets. For these reasons, a more robust procedure is needed for datasets with fewer observations.

Robust methods avoid transformation bias, as well as protecting against mis-specification of the distributional shape, by producing collective estimates for censored observations and combining these with observed data above detection limits in order to compute summary statistics. The most commonly-used robust method is called robust ROS (regression on order statistics). Robust ROS, the "MDL" procedure of Helsel and Cohn (1988), uses regression on a probability plot to estimate distributional parameters, usually in log units. Individual estimates are then predicted off the line, and retransformed back into original units. No transformation of the estimated summary statistics occurs. Similar 'robust' methods are available using a form of maximum likelihood (MLE) instead of ROS. Kroll and Stedinger (1996) demonstrated that 'robust MLE' performed somewhat better than robust ROS when estimating summary statistics. They found that the advantages cited by Helsel and Cohn (1988) for robust ROS over traditional MLE was due to the 'robust' adaptation, avoiding the transformation bias inherent with highly-skewed and/or smaller sample sizes. Kroll and Stedinger clearly shows the importance of the 'robust' adaptation, and that the order of choice for MLE and ROS methods should be: robust MLE > robust ROS >> lognormal MLE > (fully-parametric) lognormal ROS.

Essentially all of the work to date, including the 12 articles listed below, shows that ROS, MLE or K-M methods far outperform simple substitution methods such as fabricating data with one-half the detection limit. A discussion of this with an example was given in an earlier newsletter and is archived on the Practical Stats web page. Substitution can produce a signal when none actually exists, and obscure one that is truly present. It should be avoided. Its use is primarily due to the lack of knowledge of better methods.

Cohen's MLE was an early (late 50s, early 60s) method to bring maximum likelihood to those without modern computers. Cohen produced figures and tables to allow MLE to be approximated, at least for one censoring threshold (detection limit). With the availability of modern computers and MLE software there should be no reason to employ it. See the earlier Practical Stats newsletter for more on this; further discussion is also in NADA.

A reading list

Nondetects And Data Analysis includes a more detailed discussion of computing summary statistics. In the book I recommend reading 12 journal articles comparing methods for estimating summary statistics. There are many more articles available, but these dozen present the usefulness of the most common procedures, IF one understands the assumptions each article employed in doing their study. The important assumptions to look for are listed in NADA, and include whether or not 'robust' forms of methods or

their traditional counterparts are being used, the sample sizes simulated, and the distributions assumed for environmental data. Though undoubtedly my recommendations of the above three methods would not be endorsed by all of authors, the results of their work led to the recommendations I present.
The 12 articles to read are:

1. Owen, W., and T. DeRouen, 1980, Estimation of the mean for lognormal data containing zeros and left-censored values, with applications to the measurement of worker exposure to air contaminants: Biometrics 36, 707-719.

2. Gilbert, R.O. and R.R. Kinnison, 1981, Statistical methods for estimating the mean and variance from radionuclide data sets containing negative, unreported or less-than values. Health Physics 40, 377-390.

3. Gleit, A., 1985, Estimation for small normal data sets with detection limits. Environmental Science and Technology 19, 1201-1206.

4. Gilliom, R. J., and D. R. Helsel, 1986, Estimation of distributional parameters for censored trace level water quality data, 1. Estimation techniques: Water Resources Research 22, 135-146.

5. Helsel, D.R. and T. A. Cohn, 1988, Estimation of descriptive statistics for multiply censored water quality data: Water Resources Research 24, 1997-2004.

6. Shumway, R. H., A. S. Azari, and P. Johnson, 1989, Estimating mean concentrations under transformation for environmental data with detection limits: Technometrics 31, 347-356

7. Haas, C.N., and P.A. Scheff, 1990, Estimated of averages in truncated samples. Environmental Science and Technology 24, 912-919

8. Rao, S.T., J.Y. Ku, and K.S. Rao, 1991. Analysis of toxic air contaminant data containing concentrations below the limit of detection: Journ. of Air and Waste Management Assoc. 41, 442-448.

9. El-Shaarawi, A.H. and S.R. Esterby, 1992, Replacement of censored observations by a constant: an evaluation: Water Research 26, 835-844.

10. Kroll, C.N. and J.R. Stedinger, 1996, Estimation of moments and quantiles using censored data: Water Resources Research 32, 1005-1012.

11. She, N., 1997, Analyzing censored water quality data using a non-parametric approach: Journ American Water Resources Assoc. 33, 615-624.

12. Shumway, R. H., R. S. Azari, and M. Kayhanian, 2002, Statistical approaches to estimating mean water quality concentrations with detection limits: Environmental Science and Technology 36, 3345-3353.

That should keep you busy until the next newsletter!

Other articles cited:
Miesch, A., 1967, Methods of computation for estimating geochemical abundance: U.S. Geological Survey Professional Paper 574-B, 15 p.

3. Nondetects And Data Analysis textbook completed!

The upcoming book by Dennis Helsel,
Nondetects And Data Analysis: Statistics for Censored Environmental Data is at the publishers (Wiley) and being put into production. It will be available this August. A detailed outline of NADA was given in the Winter 2004 newsletter, now on the Practical Stats web site. It covers how to handle data with single and multiple detection limits, from plotting graphs to building regression equations. Look for it.

4. Next quarter's newsletter

Next topic: Lurking bias in data from labs -- the peril of recent censoring schemes.

We'd be glad to hear your comments and reactions. Email us at ask[at]practicalstats.com .

'Til next time,

Practical Stats
http://www.practicalstats.com

-- Make sense of your data