Practical Stats Newsletter for Fall, 2003

In this newsletter:
1. New course: Less Than Obvious -- December 2003
2. Correlation with censored data
3. Newsletter archive on the web
4. Topics for future newsletters

1. New Course on Handling Nondetects

Less Than Obvious, a 2-day course on the analysis of data with nondetects, will be offered Dec. 11-12 at the Univ. of California's Cooperative Extension Auditorium in Sacramento, CA. The course content of Less Than Obvious has undergone a complete remake, along with the upcoming book "Nondetects and Data Analysis" to be published by Wiley in 2004. Course content covers how to compute summary statistics, hypothesis tests, and regression models for censored data, while avoiding substituting numbers for data below detection limits. The topics for the last three newsletters are discussed in much more detail than are able to be discussed here. For more information on course content and to register for the course, go to http://www.practicalstats.com/Pages/lto.html

2. Correlation with censored data

A correlation ceofficient is one of the most commonly-used measures in statistics. It indicates the association between two variables. When one or both of those variables includes values below detection limits, how should correlation be measured?

Consider the following values of dissolved iron concentrations from Hughes and Millard (1988), collected during summers from the Brazos River, Texas.

Dissolved Iron (Y):	20	<10	<10	<10	<10	7	3	<3	<3
Time, in years (X):	1977	1978	1979	1980	1981	1982	1983	1984	1985

When one variable is time, a correlation coefficient indicates whether there is a trend in the other variable, in this case, iron concentrations. Do summer dissolved iron concentrations exhibit a trend during this period?

Published papers have reported correlation coefficients, usually Pearson's r, calculated after fabricating values for nondetects. One-half the detection limit is the value most often used. This process will give results arbitrarily dependent on the proportion of the detection limit assigned to nondetects by substitution. For example, substituting the detection limits for each nondetect in the iron dataset results in a Pearson's r of -0.89. The hypothesis test (a t-test) for determining whether $r = 0$ has a p-value of 0.001, indicating a significant trend in iron concentrations. However when

zeros are substituted for all nondetects, r equals -0.46 with a p-value of 0.216, indicating that any correlation observed is not significantly different from random noise.
No trend is indicated.

Substituting values in-between zero and the detection limit will give other results. One is no more valid than the others. Which of these conclusions, if any, is correct? It is impossible to tell, because the process of substituting numbers is flawed. The values substituted have nothing to do with what iron concentrations may have been in the bottles of water. The values are a function of the operating conditions in thelaboratory, or perhaps of the interferences in the samples. An artificial correlation may be introduced, or a real one may go unnoticed.

The alternative is to use a better correlation coefficient for censored data, one which can incorportate data censored at multiple limits. That coefficient is Kendall's tau. Kendall's tau is a nonparametric correlation coefficient commonly used in tests for trend, and one that is easily adapted for censored data. Tau is computed for a data set of n (X,Y) pairs as the number of concordant pairs of data (Nc) minus the number of discordant pairs (Nd), divided by the number of total pairs, or

Kendall's tau = (Nc-Nd) / [n(n-1)/2]

Pairs which are tied are assigned a 0. When many ties occur, Kendall (1955) proposed adjusting the denominator for the number of tied observations. This is called Kendall's tau-b. Tau-b may be more applicable than standard tau-a for the correlation of censored data.

To compute tau-b, concordant pairs Nc and discordant pairs Nd are computed for all pairs whose differences are clear. As an example, a change in Y from a <1 to a 5 after first sorting the data by increasing X is a a concordant pair, Y increasing in the same direction as X. Pairs where there are ties in X or Y, or pairs having indeterminant comparisons such as a <1 versus a <10, are considered ties. Ties do not contribute to the numerator of tau-b, and are subtracted from the number of possible pairwise comparisons in the denominator of tau-b.

Kendall's tau-b = (Nc-Nd) / sqrt[(n(n-1)/2-Ntx)*(n(n-1/2)-Nty)]

where Ntx is the number of ties in the X variable and Nty is the number of ties in the Y variable.

Consider again the multiply-censored dissolved iron data, ordered by increasing values of X. Comparisons of each observation to all subsequent observations are made, recording a + for concordant pairs and a - for discordant pairs (the sign of the slopes between two data points). Ties are assigned a zero. The pluses, minuses and zeros for all comparisons are shown below.

Dissolved Iron (Y):	20	<10	<10	<10	<10	7	3	<3	<3
sign of difference:		-	-	-	-	-	-	-	-
			0	0	0	0	0	0	0
				0	0	0	0	0	0
					0	0	0	0	0
						0	0	0	0
							-	-	-
								-	-
									0

There were 0 concordent pairs and 13 discordant pairs, so the numerator for tau-b = -13. There were no ties among the values of X, but 23 ties in the comparisons between Y observations, including the comparisons that were unclear due to censoring. Kendall's tau-b for these data is therefore

Kendall's tau-b = (0-13) / sqrt[(9(8)/2-0)*(9(8/2)-23)] = -0.60

The test for significance of Kendall's tau uses the numerator S = Nc-Nd as the numerator of the test statistic, and the standard error of S as the denominator. The equation is found in Kendall (1955), and is compared to a table of the normal distribuiton. With many ties resulting from comparisons among censored values, a tie correction is required for determining the variance of S. Software that includes tie corrections is crucial for censored data, as nondetects result in many tied comparisons. For these data, the Z test statistic is -1.50, with a two-sided p-value of 0.13. So there is insufficient evidence to prove a trend with these data.

Further detail on how to compute Kendall's tau-b and its significance test, along with software to do so, will be included in the upcoming Practical Stats course "Less Than Obvious" to be held in December. See the Practical Stats website for registration information. It will also be available in the upcoming book "Nondetects And Data Analysis" by Dennis Helsel, to be published by Wiley in 2004.

[added later] Tau-b is appropriate when all tied comparisons, including comparisons betwseen two nondetects such as a <3 to a <3, are to be ignored. Ties do not provide evidence for the null hypothesis with tau-b. On the other hand, with tau-a ties such as between two nondetects provide evidence for the null hypothesis. The scientist must decide which is more appropriate.

3. Newsletter archive on the web

Subscribers to our newsletter have asked for a place where back issues can be found and downloaded. This is now found on the Practical Stats website. There you will find the newsletters which discussed

* Is Excel and adequate statistics package?
* Why substituting one-half the detection limit for nondetects is a bad idea.
* Cohen's MLE for estimating the mean of nondetects - commonly used but outdated.

4. Your ideas welcomed

If you have ideas fir topics you would like to see discussed in future issues of this newsletter, send them to ask[at]Practicalstats.com and they may just become the subject of a future issue! All ideas welcomed.

Til next time,

Practical Stats
http://www.practicalstats.com
-- Make sense of your data