Practical Stats Newsletter for June 2011

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. Multivariate Techniques for Nondetects
3. Trailer: Statistics for Censored Environmental Data

1. Upcoming courses

Registration is open for our **Untangling Multivariate Relationships** course in Flagstaff, AZ in September. Online registration is available through the Practical Stats "Upcoming Classes" page. The registration cost is $100 lower than last year, thanks to some cost-saving efforts by holding it right after the Arizona Hydrological Society's annual symposium.
http://www.azhydrosoc.org/2011_symposium.html

Thanks AHS and USGS! There is an additional discount to people who register for the symposium – AHS will provide you a discount code when you register.

Untangling Multivariate Relationships
Turn confusion into recognizable patterns

Sept 21-22, 2011	$795 registration
USGS Water Science Center	$650 if registered for the AHS Symposium
2255 N. Gemini Dr.	
Flagstaff, AZ 86001	

NEW: We will discuss how to incorporate nondetects into multivariate procedures in this class. See below for an intro to this topic.

Also note on our Upcoming Classes page the half-day workshop "Making Sense of Nondetects" at the AHS Symposium. You don't need to register for the Symposium to attend. A link for registration is on the Upcoming Classes page.
Flagstaff is beautiful in September, near to the mountains, the Grand Canyon, etc.

You can always find our complete course listing on our "Upcoming Classes" page at
http://www.practicalstats.com/new_classes/classes.html

2. Multivariate Techniques for Nondetects

Multivariate techniques provide many more dimensions for getting into trouble when you substitute one-half or other proportions of the detection limit for nondetects. Is it really that bad? Here's what others have said:

Hopke et al. (Biometrics v57, 2001) found that substituting zero or the reporting limit produced a significant bias in results of multivariate methods. Substituting one-half or 1/square root of two times the reporting limit underestimated the variance for those variables. Both effects caused problems with later interpretations. Farnham et al. (Chemometrics and Intelligent Laboratory Systems v60, 2002) found with a simulation study that with as little as 20% censoring, slopes of principal components were not correctly computed after substitution. Substitution produced problems for cluster methods with percent censoring of 30% and higher. Their recommendation was to not use variables with 30% or more censoring. But as noted by Reimann et al. (Applied Geochemistry v17, 2002), the results of multivariate methods such as PCA and factor analysis can change radically depending on which variables are included and excluded. Excluding variables due to a specific level of censoring will potentially miss some big effects. Aruga (Analytica Chimica Acta v354, 1997) found that estimating values for nondetects from a PCA based only on the detected observations gave 'unacceptable results' with as little as less than 5 percent censoring.

So, what methods other than substitution, or deleting data, are available? In *Statistics for Censored Environmental Data*, three methods are presented. We will discuss all three in the upcoming Untangling Multivariate Relationships course in September. Below I'll focus primarily on the first.

For the first method, categorize data as below or above the (highest) reporting limit, and run procedures on the now-binary data. As an example, six DDT-related compounds were analyzed in fish. The reporting limit was at 5 ug/g. Concentrations for each of the six chemicals were given a 0 if below the limit, and a 1 if above. Then a resemblance matrix between each pair of observations was built using the simple matching coefficient, (a+d)/(a+b+c+d), where a is where both observations were a 1, d is where both were 0, and b and c are where one has a concentration above 5 and the other does not. This captures one major piece of information in censored data, the information in the proportion of observations below and above a defined threshold. Multivariate procedures operate on the resemblance matrix, and analyze the joint pattern of 0s and 1s, low and high values for each of the six chemicals.

Figure 1 is a PCA biplot of the patterns of presence and absence at 5 ug/g for the six DDT-related compounds. Note that the patterns for two groups of fish, mature and young ages, appear quite different. Also note that the pp-form of the compounds track together, increasing on the plot from left to right. The left to right gradation shows the direction of general increase to higher concentrations of the pp-forms. At the top of the plot are two sites, 11 and 30, which have high op-DDT concentrations. At the bottom are sites such as 23 where the op-forms are higher for the degradation products DDE and DDD rather than for DDT. While more detail is provided when using the other two (more precise) methods, the basic patterns in the data are all right here.

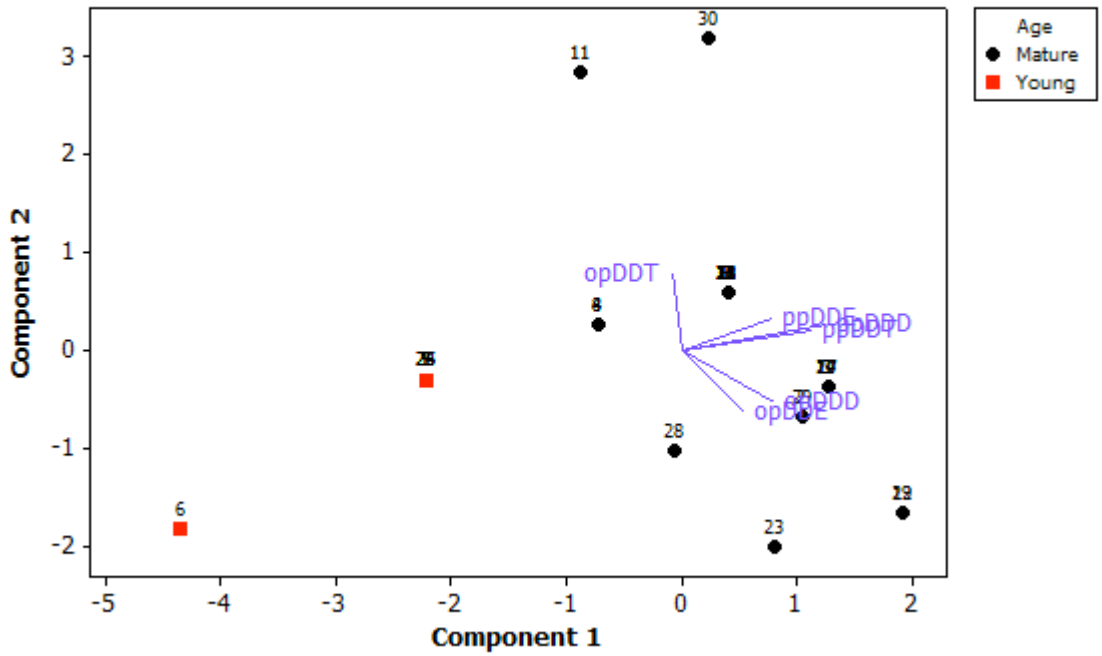

Figure 1. PCA biplot of below/above 5 ug/g concentrations of six DDT-related compounds in fish. Taken from *Statistics for Censored Environmental Data* (Helsel, 2012).

ANOSIM is a nonparametric multivariate analysis of variance, testing differences in the patterns of co-occurrence between groups. No assumption of normality is required. ANOSIM calculated on the presence/absence resemblance matrix shows a significant difference in the pattern of chemicals between the Young and Mature fish groups. The p-value is 0.001. Cluster analysis can group the sites, a factor analysis shows that the pp- and op- forms of the chemicals behave differently, and other multivariate procedures can all be computed using the presence/absence data. All without substitution.

To give an example from one of the other two methods, a multivariate extension of the Mann-Kendall test for trend is applied to the pattern of ranks of concentrations. The pattern of concentrations at the sites is tested for change over time. The test results in a p-value of 0.001, stating that a strong trend occurs in the data. This trend is pictured in Figure 2, a nonmetric multidimensional scaling (NMDS) of the ranked concentration data. Observations are colored by three time periods, showing that concentration patterns change from early to later, right to left on the plot. No values were substituted for nondetects, and there were two detection limits used for this data. The key is how to rank data with more than one detection limit. For that, stay tuned. The book should be out in the later months of 2011.

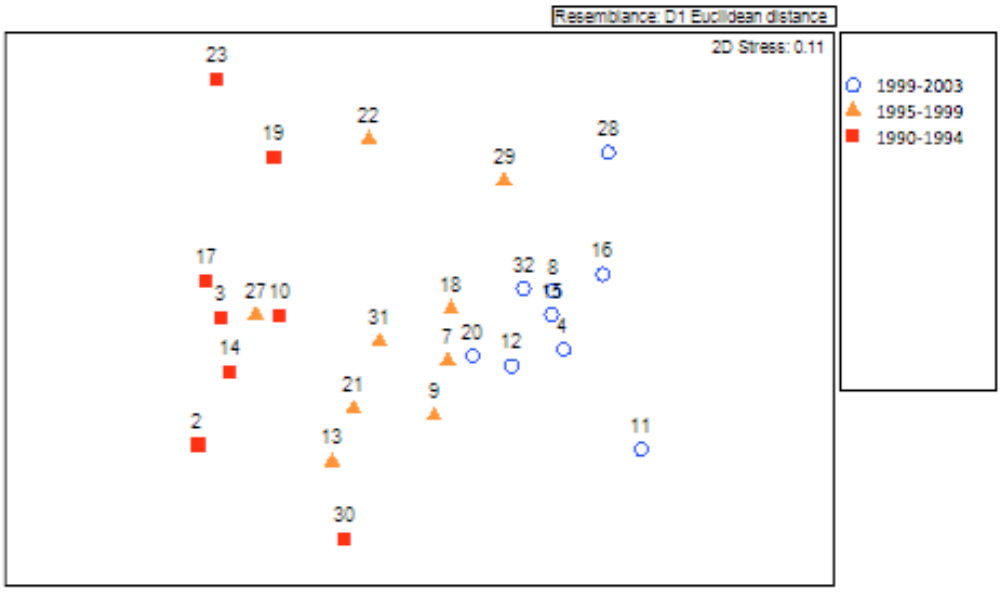

Figure 2. NMDS of locations measuring six DDT-related compounds in fish. Taken from *Statistics for Censored Environmental Data* (Helsel, 2012).

3. Trailer: Statistics for Censored Environmental Data

This newsletter is a trailer for my upcoming book, *Statistics for Censored Environmental Data*. That's the new title for the second edition of the book formerly known as *Nondetects And Data Analysis*. The new title reflects some of the new content in this second edition. It adds more emphasis on methods that consider nondetects, data below the method detection limit, as distinct from other censored data such as those pesky remarked values between the detection and quantitation limits. And the publisher much preferred the new title over my suggested change, *Pirates of the Caribbean: Curse of the Nondetects*. I was so looking forward to describing the Black Pearl Test, the procedure where nondetects just appear suddenly out of the fog. And a trend test based on that compass of Capn' Jack's. Oh well.

Topics that are new or expanded in the second edition:

- multivariate methods for censored data
- code for running procedures in R (Aaarrgghh, matey!)
- how to treat nondetects as lower than and separate from "J values" above the DL.
- summing data with nondetects
- expanded reference list to papers on censored data in a variety of disciplines
- a new introduction entitled "Invasive Data"
- and more

If you're impatient, the red-carpet premier is at the UMR class in September.

'Til next time,

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