Practical Stats Newsletter for Oct 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. Urban Legends in Environmental Statistics
3. "R you ready for some software?"

1. Upcoming courses

Applied Environmental Statistics
Statistics, down to earth

Dec. 5-9, 2011	$1495 registration
Homewood Suites	$1395 before Nov. 19, 2011
7630 Shaffer Parkway	$1295/person for 2 or more registrations
Littleton, CO 80127	

We will likely be offering our two recently-most-popular courses in March 2012 in the Atlanta GA area:

Untangling Multivariate Relationships
Turn confusion into recognizable patterns.
Relate a suite of chemical compounds to a suite of bio community structure data.

Time Series and Forecasting
Test groups, trends, etc. with frequently collected, "real-time" data.

Our **Nondetects And Data Analysis** class will reflect the upcoming 2012 release of the second edition of Dennis Helsel's book, now titled *Statistical Methods for Censored Environmental Data*. Look for that course in Spring 2012 as well.

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

Also: Upcoming workshops by Dr. Dennis Helsel.
Making Sense of Nondetects
Virginia Tech Dept. of Civil Engineering. Oct 21, 2011.

It Ain't Necessarily So: Urban Legends in Environmental Statistics
USEPA Quality Symposium, October 26, 2011, in Cincinnati, OH.

2. Urban Legends in Environmental Statistics

Seven 'urban legends' have kept statistical tools used by environmental scientists in the dark ages. These misconceptions are consistently brought by students to the four courses we teach. We spend a great deal of time in both our ***Applied Environmental Statistics*** course and our ***Nondetects And Data Analysis*** course debunking these legends. Below is a short description of why each legend is false.

1. *Parametric methods (based on a normal distribution of data) have more power than nonparametric methods.*
   This statement is true only when data EXACTLY follow a normal distribution. Field data in air, water, soils, rocks and biota never do this. If by luck or after transformation, data approximately follow a normal distribution, the power of parametric and nonparametric methods are similar. If data are skewed or contain outliers, the power of nonparametric methods is many times greater. So there is little to lose in regularly using a nonparametric procedure. A corollary: don't decide to use a parametric test until data are proven non-normal. See the Practical Stats newsletter from July 2010 for a better, more realistic flowchart for testing.

2. *t-tests determine whether one group has generally higher values than another*
   t-tests determine whether the mean of two groups are significantly different. The mean is a standardized sum. The t-test determines whether the standardized sum, or total of the two groups is significantly different. Environmental data are commonly skewed with outliers, and this can result in the means of two groups being similar while one group has higher values much more frequently than a second group. Testing whether 'one group has higher values than a second group' is a frequency question. Nonparametric tests directly measure differences in frequency of occurrence of high versus low observations. t-tests do not.

3. *R-squared is the best guide to a good regression equation.*
   There are several problems with r-squared. First, it's a function of the slope. A higher slope will result in a higher r-squared, all else being equal. The slope is rarely something under your control. Second, it always increases when a new variable is added to the regression equation, even if that variable is unrelated to the response (y) variable. Third, it's a measure of proportion of variance **in the y units used** explained by the equation. Change the y units, such as taking logs, and you can't compare directly with an equation using the untransformed y variable. An r-squared of 0.6 when y is in log units may well be explaining more of the variance in y than an equation using y itself with an r-squared of 0.7.

4. *Thirty observations are enough to apply any parametric test, such as the t-test, successfully.*
   One source of the 30 observation rule comes from the number of observations needed before the t distribution looks essentially the same as a normal distribution. This gets confused with is the number of observations needed before the Central Limit Theorem applies, and a t-test can be applied even though the data being tested do not come from a normal distribution. The actual amount of

data required before the Central Limit Theorem applies is a function of the data skewness. The more skewed the data, the more observations required. Environmental data are strongly skewed, more so than a statistician who is not used to field data expects. For skewness typical of environmental data, a number around 70 observations (per group) is a better rule of thumb.

5. *t-tests on logarithms determine whether one group's mean is higher than another.* When data are skewed, logarithms are often taken prior to performing a t-test so that transformed values are more symmetric, and so more similar to the required normal distribution. t-tests on logarithms determine whether the mean logarithm differs between the groups. When the logarithms are symmetric, the mean logarithm re-expressed into original units (the geometric mean) estimates the median. After transforming with logarithms to achieve symmetry, the t-test is testing for differences in the geometric mean or median, not the mean.

6. *Substituting one-half the reporting limit works fine if there aren't too many nondetects.*
   Substitution is not neutral. A pattern or signal is being added to the data when you substitute that may not be what you intend, and is almost certainly unlike any signal actually present in the data. The more you substitute, the more the artificial signal you've added determines the outcome of any test or model. Look at figure 1, data prior to substitution having a significant positive correlation between y and x. If some of these values were recorded only as a nondetect, and one-half the detection limit is substituted for these, the result is figure 2. You have just added a horizontal line to a large part of the data, watering down the overall correlation and slope until what was previously significant is no longer. The tragedy is that this is totally unnecessary. There are good methods for handling nondetects without substitution – see newsletters on the Practical Stats site, or Dennis Helsel's second edition book on nondetects, newly renamed as **Statistical Methods for Censored Environmental Data** (2012).

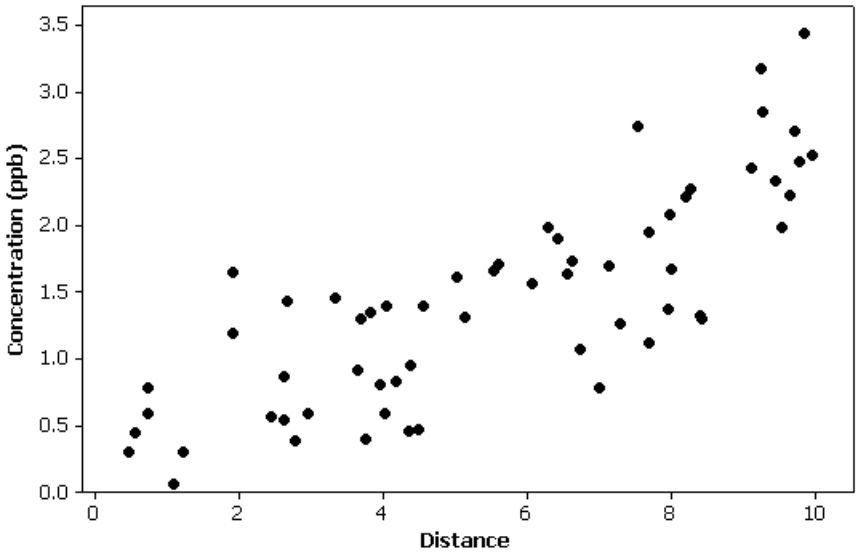

Figure 1. Original data prior to censoring. True correlation equals 0.81.

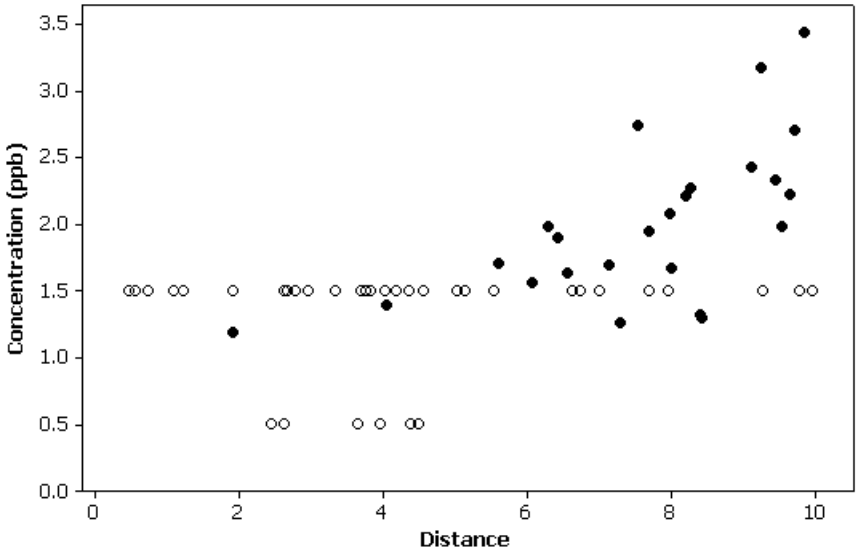

Figure 2. Data from Figure 1 after censoring at detection limits of 1 and 3 ppb and substituting ½ DL (shown as open circles). These invasive data form flat lines at one-half the detection limits, lowering the correlation to 0.55.

7. *Confidence intervals show the limits of where the next observation will likely occur.* Confidence intervals bracket the likely unknown location of the population mean, not for individual observations themselves. The mean is known to be somewhere around the center, while individual observations may not be. To predict likely locations of individual observations a prediction interval must be used, not a confidence interval.

Of course you would never believe in any of these false legends, but if you know someone who does, send them to our upcoming *Applied Environmental Statistics* course this December.

3. "R you ready for some software?"
It may not bring back Hank Williams Jr. to Monday nights, but our December *Applied Environmental Statistics* class also serves as an "Introduction to R" class. We will be using the free software R for conducting all the procedures. Replacing your current software with R may save your agency much more than the course fee in statistics software rentals alone. See our Summer 2006 and January 2010 newsletters on the capabilities of R software. All the Rocky Mountain ski areas will be open by then, so consider adding a weekend on either side to your trip. The host hotel, the Homewood Suites in Littleton is less than an hour from the major ski areas, yet in its SW Denver location is outside of the much higher-priced ski alley.

'Til next time,

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