Practical Stats Newsletter for August 2014

Subscribe and Unsubscribe: http://practicalstats.com/news
Archive of past newsletters http://www.practicalstats.com/news/bydate.html

In this newsletter:

1. Upcoming Training

In-person courses:
Untangling Multivariate Relationships
Oct. 21-22, 2014 $895 through Oct 5, $995 after.
Austin, TX
http://www.practicalstats.com/training/umr/

Nondetects And Data Analysis
Oct. 23-24, 2014 $895 through Oct 5, $995 after.
Austin, TX
http://www.practicalstats.com/training/nada/

Time Series Methods (for frequently collected, "real-time" data)
Oct. 28-29, 2014 $895
Littleton, Colorado
http://www.practicalstats.com/training/timeseries/

Statistics for Contaminated Sites
Nov 14, 2014 (cost and details to come)
Vancouver, BC
http://www.practicalstats.com/training/contam/

To register and for more information on all of our courses and webinars, see our Training page at http://practicalstats.com/training/

Upcoming talks for Dennis Helsel:
Chemometrics for Data with Nondetects. 6th International Chemometrics Research Meeting, Nijmegen, Netherlands. Sept 14-18, 2014.

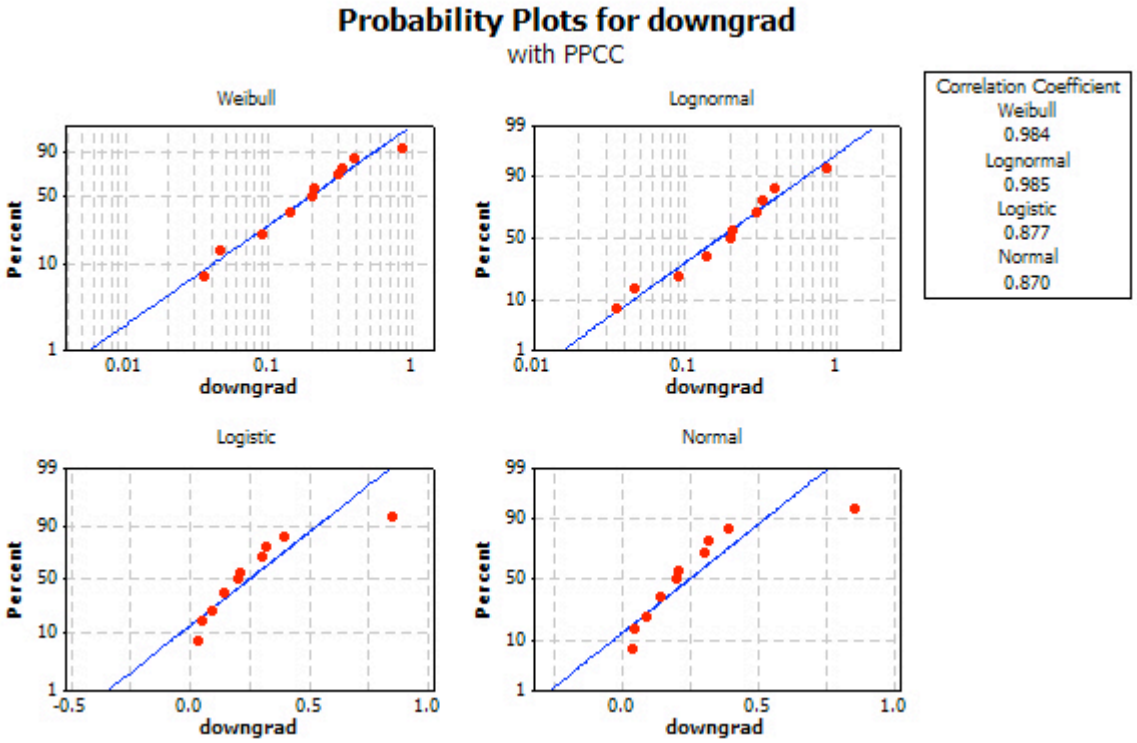
Obtaining Confident Statistical Test Results: Choosing parametric vs nonparametric tests is obsolete. SETAC 35th Annual Meeting, Vancouver, BC, Nov 9-13, 2014.

2. Testing for a Normal and Other Distributions

Testing for a normal distribution prior to deciding which test for group differences to use is becoming less and less necessary. Powerful tests for differences in group means are now available that do not require an assumption of normality. We highlight these in our Applied Environmental Statistics course, and I'll be speaking about them at the SETAC conference coming up this November.

However, there are situations when it helps to fit a distribution to data. Three are when computing confidence, prediction, and tolerance intervals, particularly for smaller sample sizes. A distributional 'model' allows the scientist to project higher percentiles even with insufficient data to compute them with only the data at hand. For example, the 95th percentile is the value exceeded only 5 out of 100 (or 1 out of 20) times. The fewest data required to estimate this percentile without using a model is 19, where the maximum of the 19 observations becomes the first available estimate of this high percentile. If you have 18 or fewer observations your only recourse is to fit a distribution to the data.

Three tests for distributional fit are today recognized as being powerful -- power is the ability to see differences from the assumed distribution when that distribution is a poor fit. Tests for distributional fit assume the data follow the distribution, and conclude the fit is poor when the p-value computed from the data drops below 0.05 (or your chosen alpha). The three more-powerful tests are the Anderson-Darling (AD) test, the Shapiro-Wilk test, and the Probability Plot Correlation Coefficient (PPCC) test. The PPCC test computes Pearson's r between the x and y axes of a distribution's probability plot. Here is Minitab's report of how well 4 distributions fit a 13-observation dataset of molybdenum concentrations:



The PPCC of the Weibull and lognormal distributions are closest to 1, and of the distributions examined fit the data most closely. Using a well-fitting distribution is a useful approach for estimating high percentiles or other statistics for small data sets. The p-value for the Shapiro-Wilk test of normality is 0.006, soundly demonstrating that the data do not follow a normal distribution.

A second class of test procedures have much lower power when testing for distributional fit. These are the Kolmogorov-Smirnov (KS) test, the Lilliefors test and the chi-square test. All three are occasionally found in textbooks, and sometimes in software, but cannot see differences from a distribution as readily as do the first class of procedures. The KS test assumes the user knows the true mean and standard deviation, which is rarely if ever true in environmental studies. The Lilliefors test modifies the KS test to use the mean and standard deviation computed from the data; otherwise it is the same test. Below is the output for the chi-square and Lilliefors tests of normality using the nortest package of R.

```
> pearson.test(MOLY)

        Pearson chi-square normality test
data:  MOLY
P = 4.0769, p-value = 0.2533

> lillie.test(MOLY)

        Lilliefors (Kolmogorov-Smirnov) normality test
data:  MOLY
D = 0.2115, p-value = 0.1152
```

Note that neither are able to find a difference from the assumed normal distribution, while the Shapiro-Wilk test clearly did. Thode (*Testing for Normality*, CRC Press, 2002) states that due to their lack of power, these latter three tests should be avoided when checking for a normal distribution.

3. HHARG

Rumor has it that *Statistical Methods for Water Resources* (USGS, 2002), otherwise known as Helsel and Hirsch, is being updated for the modern era. Three new authors (Stacey Archfield, Karen Ryberg and Ed Gilroy) are busily working with the two old authors to provide new twists on methods, and examples using R statistical software. While we can't confirm the rumor, sources tell us copies may be hidden in various public places around North America sometime in 2015, and clues to their locations leaked through social media. @hiddenstats

'Til next time,

Practical Stats
-- Make sense of your data