Practical Stats Newsletter for May 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/

Webinars:
Our fall webinars will be announced by Twitter, on our web and Facebook pages, and in the July newsletter.

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

2. The Meaning of a Mean

This is an update of an article I wrote thirty years ago this month for a US Geological Survey internal magazine (the WRD Bulletin). It focused on whether or not to use a mean when describing and testing pH. The topic is still relevant, as two of our Top

Twelve Tips (see our Blog page or the March 2014 newsletter) involve the same issues. As Yogi Berra first said, “Déjà vu all over again”.

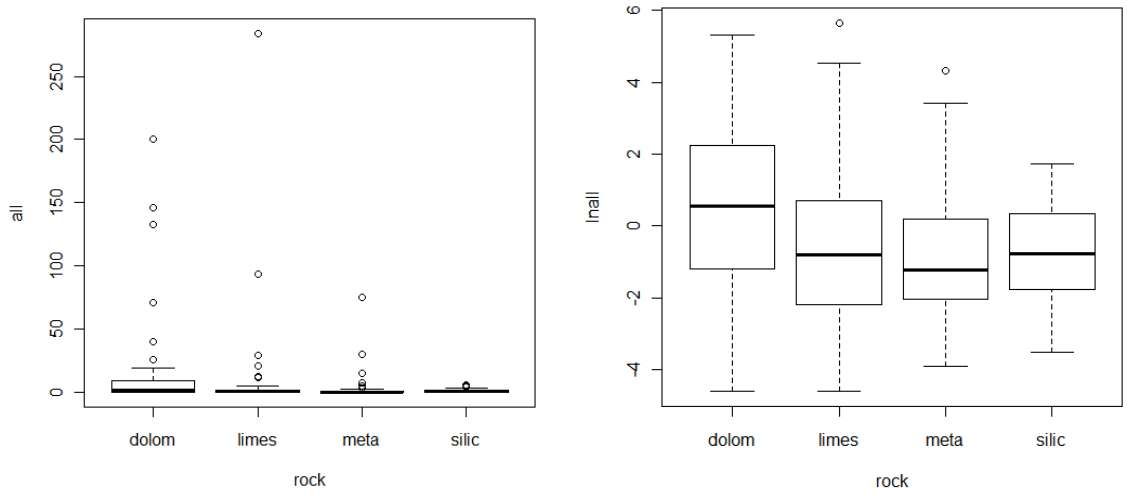
In general, and in regard to pH in particular….
What is the correct numerical summary to use, mean or median? Which statistic should be used when testing for group differences?
The answer should depend solely on your objectives.

If your goal is to measure the mass of contaminant (or acidity) transported down a river for a given year or in a specific area of an aquifer, you sum each day’s or location’s contribution to obtain the total. Dividing the total by the number summed is the mean -- the mean is a standardized total. The mean is a good choice for expressing the cumulative mass on a per-measurement basis. When your interest is in mass, volume, cumulative exposure or other measures to be summed, a mean is the appropriate summary statistic.

Alternatively, if your goal is to express the typical concentration, the median is a better choice than the mean. A median is resistant to the effect of unusual values. When nine of ten locations have low concentrations but the tenth is much higher, the median is relatively unaffected by the one high value and looks much like the concentrations in the other nine. The mean would be pulled up toward the high value, sometimes being higher than all of the other nine observations. When the interest is in a representative value, the median is the appropriate summary statistic. Transforming a median pH back to H+ units estimates the median H+. Medians move across scales. However the mean pH does not estimate the mean H+ when transformed back – it becomes a geometric mean H+, a second estimate of median H+. Means are unit-specific.

If your interest is in testing the difference between groups in totals of all values, test differences in mean values (this is rarely the case in environmental science). The question “does one group generally have higher values than another?” is a frequency question -- do higher values occur more frequently in one group? Nonparametric methods answer this exact question. Do not decide which type of test to use based on whether data follow a normal distribution. Permutation tests discern differences in means, and nonparametric tests discern differences in frequencies, for all shapes of normal and non-normal data. Decide which type of test to use based on the objectives of your study.

pH is the negative log10 of hydrogen ion activity (concentration), and was developed because the response of pH electrodes is linear with the log of activity. Logarithms have long been a logical scale for environmental data due to the multiplicative effects and skewed distributions of natural systems, and the percent rather than absolute errors often shown by lab equipment. But what does working in log units do to the interpretation of the result? As an example, specific capacity, a standardized measure of yields of water from wells, was measured in hundreds of wells across the Appalachian region of the US in a USGS report of the late 1980s. The data come from four rock types, and there was strong interest in learning if well yields differed between the four rock units. The plot on the left shows the original data, and is analogous to H+ concentration data. The plot on the right shows log-transformed well yields, analogous to measurements on a pH scale.



Analysis of variance (ANOVA) is a parametric test requiring data within groups to follow a normal distribution. For the non-normal original data the p-value is above 0.05, so group means were not found to be different. Is the non-normality of data causing a loss of power, pushing up the p-value, even with 50 observations in each group? ANOVA on the logarithms (right-hand plot) gives a p-value of 0.007. This tests the difference in the means of the logarithms, which transforms back into geometric means (medians) of well yields. Medians or typical values differ, while the means cannot be seen as different by ANOVA. Similarly, an ANOVA on pH tests differences in geometric means (medians) of H+.

A nonparametric test of median pH also tests median H+ concentrations. If a test of mean H+ for skewed H+ data were needed, a permutation test will have more power than a parametric test such as ANOVA. The permutation p-value on the untransformed specific capacity data above is 0.04 despite the data's skewness, stating that group means do differ. The permutation test can see this difference in mean specific capacity while ANOVA could not, due to the non-normal shape of data.

Nonparametric tests of medians and permutation tests of means simplify the hypothesis test process, whether done in pH (log) units or in H+ concentration units, as they accept data of any shape. Parametric tests on skewed H+ data will have low power. Remember that regardless of which test for means you use, parametric or permutation, testing means in pH (log) units does not test means in the untransformed concentration units, but tests geometric means (medians) in those units instead.

3. Talks, Tweets and Videos
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.

Webinars on-demand. Yes they are still coming. You will be able to sign up and view the videos any time of the day, at your convenience. This will also let those outside the North American time zones view the material and interact with the instructor by email. Registration will be through our Training page. Hopefully, all will be set up in June. We'll tweet when ready, and put the info on our Facebook and web pages.

'Til next time,

Practical Stats
-- Make sense of your data