Practical Stats Newsletter for May 2015

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

In this newsletter:
1. Upcoming 2015 Training
2. The Power of Permutation Tests
3. Software for Permutation Tests

1. Upcoming 2015 Training

In-person courses:

Permutation Tests
Aug. 24-25, 2014 $995 through Aug 9, $1095 after.
Altamonte Springs (Orlando area), Florida
Permutation test procedures replace parametric tests like t-tests and ANOVA. Learn about these new, important methods for environmental statistics.

Applied Environmental Statistics
"Statistics down to earth". A complete survey of statistical methods for environmental data, as well as an introduction to using R statistical software.
Sept. 14-18, 2015. $1495 through August 30, $1595 after.
Lynnwood (Seattle area), Washington

Webinars:

* Permutation Tests: Never Worry About a Normal Distribution Again!
Monday June 22, 2015. 1:30-3:00 PM Eastern, 10:30-12 noon Pacific, $250
Introduction to the new permutation test procedures that replace t-tests and ANOVA. See the benefits of the new, important methods for environmental statistics. (Our two-day class on permutation tests, above, will provide much more detail and hands-on use of software in class.)
To register and for more information on all of our courses and webinars, see our Training page at http://practicalstats.com/training/

2. The Power of Permutation Tests

Consider four locations where the concentration of a contaminant has been measured. We would like to know whether the mean concentration differs between the sites, as the mean is specified in a regulation for the contaminant. Boxplots of the data are shown in Figure 1, with means shown as open circles with crosshairs inside. Site 1 is a background site. The other sites appear to be somewhat different in concentration than background. The data are skewed – concentrations do not follow a normal distribution.

An analysis of variance (ANOVA) produces a p-value of 0.07, not sufficient evidence to declare that the four means differ from one another. We can't say from the ANOVA that the mean concentration at sites 2 to 4 are anything other than background. However, we remember from a Practical Stats training class that ANOVA is a parametric test that can have low power (higher p-values than it should) when data do not follow normal distributions. What can we do about this?

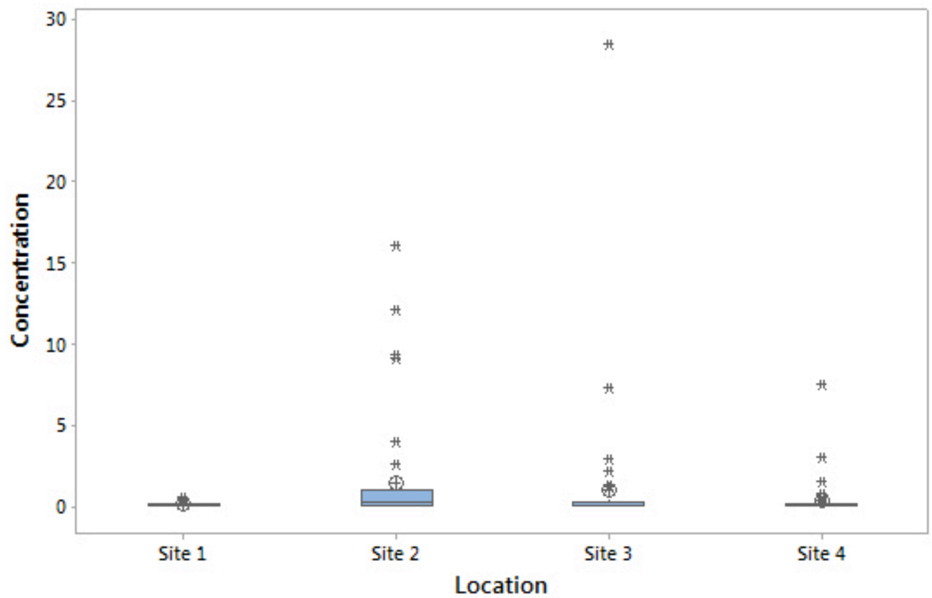

Figure 1. Boxplots of the data and mean values

In the past people often took logarithms of the data in this situation. See Figure 2. Logarithms of skewed data are often closer to a normal distribution than the original units. The ANOVA on logarithms has a p-value of 0.007. Mean logarithms do differ.

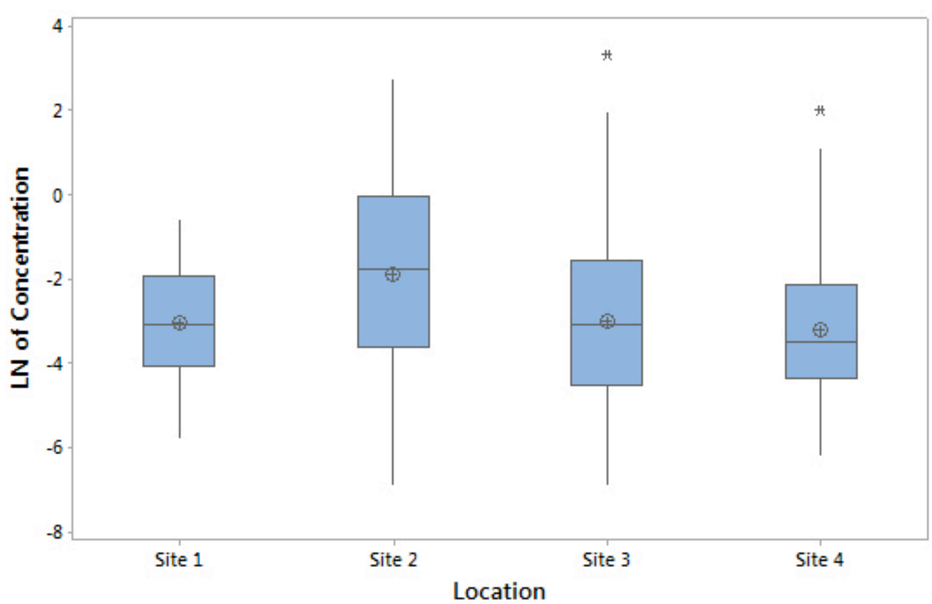

Figure 2. Logarithms of concentrations with mean logarithms as circles

The mean logarithm, re-transformed back to original concentration units, is the geometric mean of each group's data. The geometric mean is an estimate of the median of the data, not of the mean. After taking logs we are no longer testing for differences in the means. So this doesn't help us in distinguishing mean concentrations.

Some people perform a nonparametric alternative to ANOVA, the Kruskal-Wallis (KW) test, when data are skewed. KW has a p-value of 0.009, very similar to the results of an ANOVA on logs. That is because a KW test is also testing for differences in medians, not means. It is an excellent test of "does one group exhibit higher concentrations than another?", but it also is not a test for differences between means.

Until recently, testing differences in means for the skewed data common to environmental studies typically required 70 or more observations. With sufficient data, the Central Limit Theorem states that p-values resulting from parametric tests such as ANOVA will be correct even when the data don't follow the required distribution. However, environmental studies often do not have the luxury of that much data. Is there no good solution?

Permutation tests solve the problem of testing means for moderate sized, skewed data sets. No assumption of any distribution is required to obtain a p-value. Instead, several thousand permutations of the data are used to compute several thousand test statistics, and the proportion of test statistics exceeding the one observed (for ANOVA, the F statistic) is the p-value for the test. For our example, the permutation p-value is 0.043 (Figure 3).

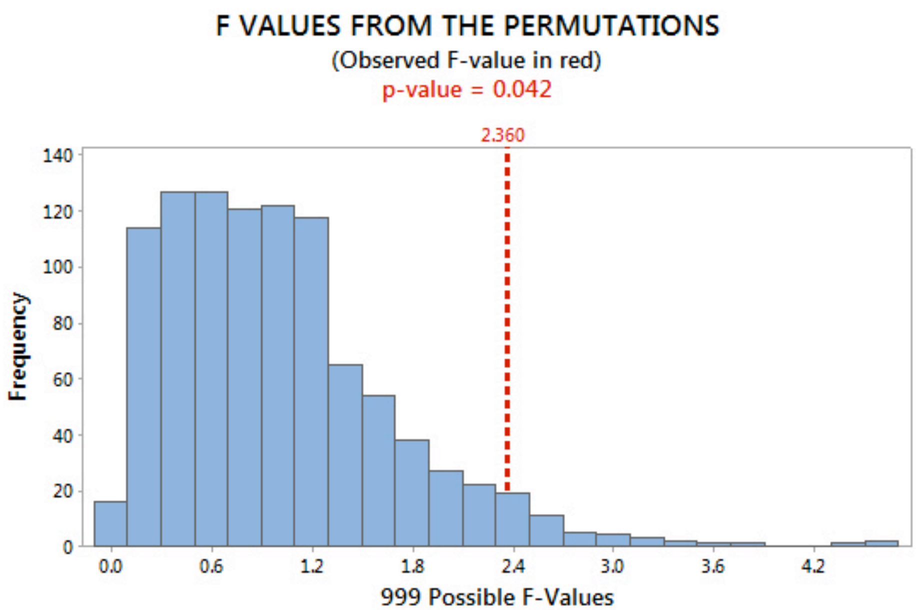

Figure 3. Histogram of permutation F-statistics. 4.2% equal or exceed the observed ANOVA test statistic of 2.36.

The difference between the ANOVA p-value of 0.073 and the permutation p-value on the same data of 0.042 is the loss of power caused by the ANOVA's assumption of a normal distribution when the data do not follow that distribution. Using the permutation test, we demonstrate that there is a difference between the four mean concentrations. The outcome differs from ANOVA without any additional cost of data collection. The permutation test simply makes more efficient use of the information. Shouldn't you increase the efficiency of your data analysis?

3. Software for Permutation Tests

Permutation tests are relatively new – it takes about 20 years for commercial software companies to decide whether or not to include a new procedure into their lineup – will it help sell a new version of the software? Currently SAS, Stata and SPSS include permutation tests. Mid-level packages such as Minitab, Statistica and others do not yet include them. The free R statistics system has the largest collection of permutation procedures. You'll learn more about R's implementation of perm tests in our one-week Applied Environmental Statistics (AES) course this September, or in the 2-day Permutation Test course in August. If you have taken our AES course prior to 2009, the two-day Permutation Test class will be the most cost-effective way to get up to speed with these new methods.

'Til next time,

Practical Stats
-- Make sense of your data