Practical Stats Newsletter for July 2016

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

In this newsletter:
A. Upcoming Webinars and Talks
B. Nonparametric Two-factor ANOVA
C. The Third Section of Our Newsletters

A. Upcoming Webinars and Talks

Our webinar "Seven Perilous Errors in Environmental Statistics" is currently available for free at our Online Training Center (http://practicalstats.teachable.com).

Our "Permutation Test and Bootstrapping" course will be at the Online Training Center within two weeks, on demand at any time you want to start.

Our "Applied Environmental Statistics" comprehensive survey course will be at the Online Training Center this fall, on demand at any time you want to start.

We are also glad to come to your site and teach any of the six courses we offer in-person. See http://practicalstats.com/training/ for details.

B. Nonparametric Two-factor ANOVA

Our Applied Environmental Statistics course, which will be offered online this fall, covers a variety of group hypothesis tests. A familiar design is two-factor ANOVA, evaluating the effects of two influence factors on the mean, for example a location and a seasonal factor. With three locations and four seasons there are 12 groups of data, each of which should follow a normal distribution and all of which should have the same variance in order for p-values of ANOVA to be correct. For environmental studies, those assumptions are rarely met. The effect of violating these assumptions is the same as for one-way ANOVA, a loss of power leading to higher p-values and failure to find significant differences that are there.

A nonparametric two-factor test that does not require normality or equal variance is the BDM test. It determines whether the cumulative distribution functions (cdfs, the percentiles of data) have dissimilar patterns due to one or more factors. Applied here to the two-factor layout, it is more powerful than the 'two-way ANOVA on ranks' method of Conover and Iman (1981) that was for many years the nonparametric two-factor test most familiar to environmental scientists. In their simulation study, Brunner et al. (1979) found that when all assumptions of ANOVA were met, the BDM test had nearly the same power to detect factor effects as did classic ANOVA. With heteroscedasticity and/or non-normality, BDM has greater power than ANOVA. They summarize by saying that BDM "represents a highly accurate and powerful tool for nonparametric inference in higher-way layouts."

The null hypothesis of the BDM test is that there are no changes in the distribution function of data due to factor A, factor B, or to an interaction. After some interesting yet "computationally simple" matrix algebra on ranks (distribution percentiles), a shift in the distribution function due to factor A or B will cause an increase in the F-test statistic, and rejection of the null hypothesis. A shift beyond what is due to either factor in one or more individual group cells will cause the test for interaction to be significant. The procedure is quite analogous to analysis of variance (ANOVA) with the difference that shifts in group distribution functions are being evaluated, rather than shifts in group means. In that sense it is like a two-factor version of the Kruskal-Wallis test. The BDM test in implemented in the asbio package of R with command BDM.2way.

As an example, iron concentrations were measured at low flow in numerous small streams in the coal-producing areas of eastern Ohio (Helsel, 1983). One factor is mining history (unmined areas, abandoned mines, or reclaimed mines) and the second factor is rocktype (sandstone or limestone). Are iron concentrations at low flow influenced by upstream mining history, by the underlying rock type, or by both?

Boxplots for total iron concentrations are shown in figure 1. Note the skewness, as well as the differences in variance as depicted by differing box heights.

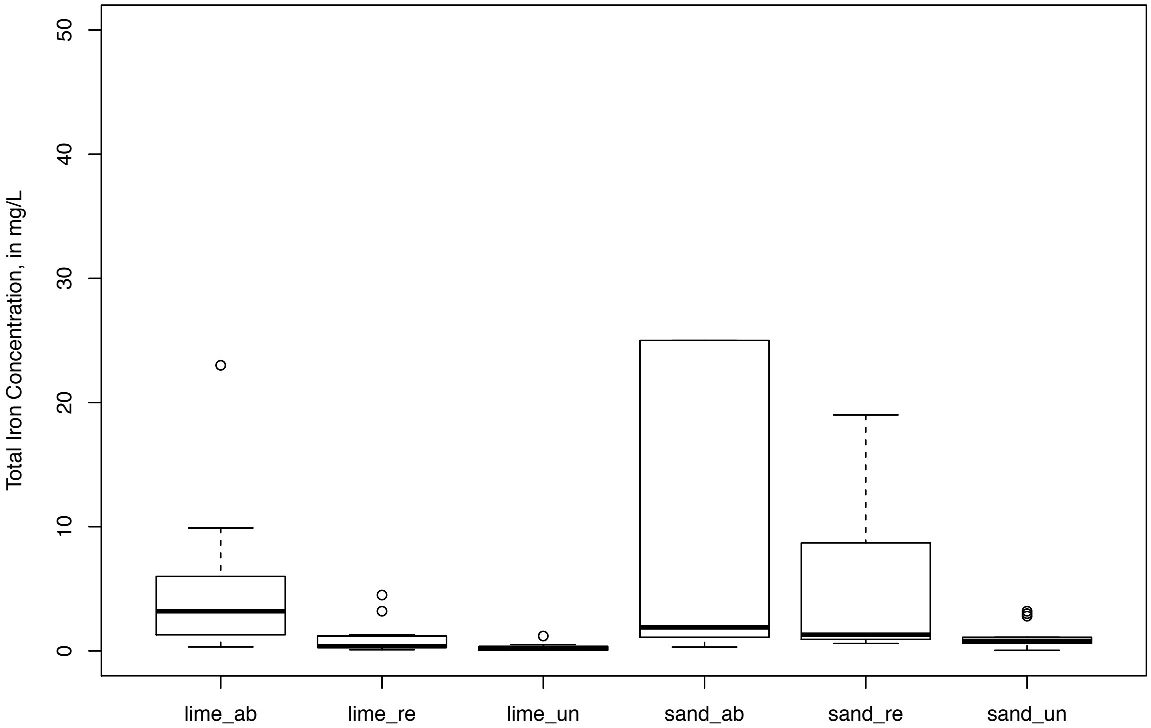

Figure 1. Boxplots of iron concentrations at low flow from Helsel (1983). Three outliers above 50 mg/L are not shown.

First we run ANOVA and find no difference attributed to either factor or their interaction:

```
> summary(aov(fe~mining*rocktype)
                Df Sum Sq Mean Sq F value Pr(>F)
mining           2  32282   16141   2.493 0.0898
rocktype         1  15411   15411   2.380 0.1273
mining:rocktype  2  25869   12934   1.997 0.1431
Residuals       72 466239    6476
```

A Shapiro-Wilk test finds that residuals are non-normal. Could this "not significant" ANOVA be due in part to violation of assumptions? To determine this, run the BDM test:

```
> BDM.2way(fe,mining,rocktype)
Two way Brunner-Dette-Munk test
           df1      df2        F*     P(F > F*)
X1    1.981511 64.36822 17.740921 7.885850e-07
X2    1.000000 64.36822 13.375242 5.152032e-04
X1:X2 1.981511 64.36822  3.709541 3.023646e-02
```

X1 is the first factor listed in the command (mining history), and X2 is the second factor listed (rocktype). The interaction term is X1:X2. While parametric ANOVA was unable to find significance for either factor or the interaction, all three tests are significant using the nonparametric BDM test.

We discuss other approaches to two-factor analysis in the AES class, including permutation methods. The bottom line is that environmental data often violate the assumptions of classic parametric methods to the level that they are unable to detect important patterns and relationships. But there are other, better methods available.

References:

Conover and Iman, 1981. Rank transformation as a bridge between parametric and nonparametric statistics: *The American Statistician 35*, p.124-129

Brunner, Dette and Munk, 1997, Box-Type Approximations in Nonparametric Factorial Designs. Journ. Am. Stat. Assoc. 92, p.1494-1502

C. The Third Section of Our Newsletters

While section B of our newsletters provides detailed information on environmental statistics, you never know what you'll find in section C. Last time we gave you a link to an online comic, xkcd.com, which we find pretty humorous (though nerdy). This month, how about two useful links to statistical information online?

1. PAST. Free software that does all the basics and many multivariate methods.
http://folk.uio.no/ohammer/past/index.html
Runs on macintosh and Windows computers.
2. The Handbook of Biological Statistics by John McDonald of the Univ. of Delaware.
http://www.biostathandbook.com/index.html
A good online source for parametric and basic stats.

Perhaps these won't confuse you as much as last time. But watch out for next time.

'Til next time,

Practical Stats
-- Make sense of your data