Practical Stats Newsletter for March 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. What are the Seven Perilous Errors?
C. No, there wasn't a January newsletter!

A. Upcoming Webinars and Talks
Our webinar "Seven Perilous Errors in Environmental Statistics" is now available for free, online. I've made it the topic of this newsletter as well, to introduce you to my thoughts on seven common and damaging errors made by environmental scientists.

Webinars:
Seven Perilous Errors in Environmental Statistics
Now available for viewing at http://practicalstats.teachable.com
or through our Training page: http://practicalstats.com/training/

Online Courses:
Permutation Tests
Never Worry about a Normal Distribution Again!
$550 until April 25th, $600 after. Begins May 10th.
See http://practicalstats.com/training/

B. What are the Seven Perilous Errors?
Our January 26th webinar, now available for viewing online, reported seven serious errors commonly made by environmental scientists when interpreting their data.

Error 1. A significant p-value tells you all you need to know

- Understand what a p-value is: the probability of a false signal, such as a false difference between groups or a false determination of trend.
- A small p-value in regression doesn't mean that you have the best regression model possible. If you haven't checked, you might even be fitting a straight line to data with a curved pattern. Plot the data before you ever look at a number on the output!
- Statistical significance is NOT the same as usefulness. Always look at the magnitude of the difference after a statistical test indicates it is non-zero. Even though significant, that difference may be small enough to be of no practical interest.

Error 2. Testing for a normal distribution to decide whether employ a parametric or nonparametric test

- Field data rarely if ever follow a normal distribution

- For skewed data with outliers, nonparametric methods have a large power advantage (will find more differences when they are there) over parametric tests
- Most importantly, tests on means (parametric tests) answer different questions than tests on typical patterns (nonparametric tests). Decide which question you want answered, and use that type of test. A test for trend, for example, is a frequency test – do concentrations generally increase over time? Nonparametric tests answer this question directly. A parametric test answers it only under specific conditions.

Error 3. Using t-tests and ANOVA with small data sets

- ANOVA and t-tests are only accurate when data follow a normal distribution. Or with 70+ observations per group (my general rule of thumb for when the Central Limit Distribution applies with data as skewed as that of environmental science).
- When a test for difference in means is the objective, a nonparametric test doesn't help. It tests for differences in percentiles (frequencies or typical patterns). Taking logs doesn't help (see next error).
- Instead, test for differences in means using permutation tests. Older normal-theory tests were only approximations (according to Karl Pearson, one of their developers). Those approximate methods are no longer necessary. Look for software that includes permutation test methods.

Error 4. Testing logarithms to look for differences in means

- Testing differences between group means in log units (with ANOVA or t-tests) does NOT test for differences in means in original units. Instead it tests whether the geometric means differ between groups.
- The geometric mean is an estimate of the median, not the mean, of the distribution. These are two different things.
- If you want to test differences between means of non-normal data, use a permutation test. If you want to test for "does one group show higher values than the second?" instead, that is directly answered using a nonparametric test.

Error 5. Using only r-squared to find the best regression equation

- r^2 is "in units of" the y-variable. Changing the units (taking the log of y, etc) puts the statistic in a different set of units. Comparing it to the r^2 in original units.
- r^2 depends on the slope. Higher slope = higher r^2, all else being equal. We have no control over slopes
- Better, more modern numerical criteria are available for determining the best regression equation

Error 6. Using outlier tests to find and delete 'bad' data

- Outlier tests cannot tell you whether data are 'wrong', only that they aren't likely to have come from a normal distribution
- Environmental field data (water, air, soils, rock, biota) are usually skewed distributions, not originating from the normal distribution. There are at least three reasons why outliers are to be expected.

- Albert Einstein was an outlier

Error 7. Substituting one-half the detection limit for nondetects

- This is not a good procedure! There are better methods.
- Substitution produces an invasive, false signal (or false no-signal). You could find differences that are not really there, or not find differences that are.
- One of the simplest, better methods is to use a nonparametric test. This will rank all nondetects as the lowest values, tied with one another, which is exactly what you know about the data. If there are multiple detection limits, you'll need to censor all data below the highest detection limit as a <highest DL and then run the test. Even this will be better than substitution followed by something like a t-test or ANOVA.

There is more information in the online webinar. Take a look for yourself. The first six errors are also fully discussed in our Applied Environmental Statistics course, which will be offered online this coming fall. The seventh is covered along with much more about handling nondetects in our Nondetects And Data Analysis course, which will also be offered online later this year. Availability of all courses will be announced in this newsletter, and on our Training web page.

C. No, there wasn't a January newsletter!

You didn't miss it, it didn't happen. It takes time to get content online in a way that provides solid training. We've started by placing the "Seven Perilous Errors" online now, and conducting our first online course on Permutation Tests this May.

'Til next time,

Practical Stats
-- Make sense of your data