Practical Stats Newsletter for January 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
2. Nonparametric Tests for Censored Data With "DL to RL" (or Remarked) Values
3. Our Top Twelve Tips

1. Upcoming Training

In-person courses (see our Training page to register):
Applied Environmental Statistics
"Statistics, Down to Earth"
May 5-9, 2014 $1395 through April 13, $1495 after
Indianapolis, IN
Applied Environmental Statistics covers statistical methods for analysis of air, water, soils, and bio data. It includes how to build good regression models, a myriad of hypothesis tests including the newer permutation tests, and trend analysis. It enables you to make sense of your data. A full course outline is on our website.

Webinars (see our Training page to register):
1. Statistics for Managers
March 20, 2014 11am-noon Mountain, 1-2 pm Eastern. Free.
Why would my employees need to know more than that one-semester course they took 5-10 years ago in college? Think about what's changed over the last 10 years – tablets, ease of access to wi-fi, Facebook and Twitter…… Statistics has changed a lot as well. We'll hit the high points of

- flexible tests with few requirements for validity,
- free comprehensive statistics software,
- new and better methods of finding the best regression line,
- why "has there been a change in concentration?" has very little to do with a mean,
- and much more.

All without jargon or equations, or selling software. Just how this might benefit your firm's/agency's products. A Q&A time will follow.

Pass this along to people in your office who are tasked with 'the big picture'.

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

2. Nonparametric Tests for Censored Data With “DL to RL” (or remarked) Values
How can I incorporate data recorded as between the detection and quantitation/reporting limits? Answer: Express data as an interval (low, high) of the possible values each observation may take. For example, values below a detection limit of 1 are expressed as (0, 1). Values between the detection limit and a quantitation limit of 3 are expressed as (1, 3) rather than remarked as 2.7J. Detected values possess the same number for both low and high values, so that a detected 5 is (5, 5). Data expressed in this way are “interval-censored”.

Fay and Shaw (2010) wrote the contributed package "interval" for R to compute interval-censored nonparametric tests for cancer data. Peto and Peto (1972) described these tests, but software to accomplish them was not easily available until this contributed R package was released. These procedures extend the rank-sum type score tests to data expressed in the interval endpoints format, allowing environmental scientists to test datasets with 'detected but not quantified' data directly. Interestingly, in their paper Fay and Shaw refer to an evaluation by Law and Brookmeyer (1992) on substituting one-half the interval width for interval-censored data -- the equivalent of substituting one-half the reporting limit when the lower endpoint is zero. Not surprisingly, it didn't work very well. Hence the need for this software to avoid substitution when a disease occurs somewhere between time A and time B. Environmental scientists can use the same procedures to test concentrations that fall between A and B ug/L. In fact, one of the first papers discussing interval-censored tests had applied them to left-censored chemical data, concentrations of PCBs built up in the human body (Self and Grossman, 1986). Full citations are found in my textbook *Statistics for Censored Environmental Data using Minitab and R* (2012), from which this newsletter is extracted.

How does the test work? Linear rank tests estimate the survival curve (percentiles) and determine if this differs between the groups of data. Estimates of the probability of exceeding each cutpoint (detection limits and detected observations) for each group are compared to the overall probability of exceeding those cutpoints for all data together if the null hypothesis is true and there is no difference among the groups. A score is computed summarizing the differences between the within-group exceedance and the overall exceedance probabilities. One option in the interval package is to use the "rho=1" score statistic, producing an interval-censored analog to Wilcoxon-style tests such as the rank-sum or Mann-Whitney test.

To illustrate, cadmium concentrations were measured on stream sediments from two different geologic regions in the Rocky Mountains (a US Geological Survey dataset that comes with the NADA for R package). The original data are re-expressed so that values between a detection limit of 0.4 and a reporting limit of 0.6 are seen as (0.4, 0.6). Values below the detection limit are seen as (0, 0.4). See the textbook for the meaning of indicators Rt and Lt. The low end of the interval is the Cdlo column; the high end is the CD column:

```
> Cd2
     CD   REGION LT.1 Cdlo    Rt   Lt
1  81.3 S RKY MT    0 81.3  TRUE TRUE
2   3.5 S RKY MT    0  3.5  TRUE TRUE
```

```
3   4.6 S RKY MT    0  4.6  TRUE TRUE
4   0.6 S RKY MT    0  0.4 FALSE TRUE
5   2.9 S RKY MT    0  2.9  TRUE TRUE
6   3.0 S RKY MT    0  3.0  TRUE TRUE
7   4.9 S RKY MT    0  4.9  TRUE TRUE
8   0.6 S RKY MT    0  0.4 FALSE TRUE
9   3.4 S RKY MT    0  3.4  TRUE TRUE
10  0.4 COLO PLT    0  0.0 FALSE TRUE
11  0.8 COLO PLT    0  0.8  TRUE TRUE
12  0.4 COLO PLT    1  0.0 FALSE TRUE
13  0.4 COLO PLT    0  0.0 FALSE TRUE
14  0.4 COLO PLT    0  0.0 FALSE TRUE
15  0.4 COLO PLT    1  0.0 FALSE TRUE
16  1.4 COLO PLT    0  1.4  TRUE TRUE
17  0.6 COLO PLT    1  0.4 FALSE TRUE
18  0.7 COLO PLT    0  0.7  TRUE TRUE
19  0.4 S RKY MT    1  0.0 FALSE TRUE
```

The test for difference between regions is run using the ictest command of the interval package:

```
> cd2test=ictest(Cdlo,CD,REGION,rho=1,Lin=Lt,Rin=Rt,exact=TRUE)
> cd2test
        Exact Wilcoxon two-sample test (permutation form)
data:  {Cdlo,CD} by REGION
p-value = 0.00747
alternative hypothesis: survival distributions not equal
```

The exact test uses either all possible comparisons to compute a p-value (this data set is sufficiently small to do that) or a random subset of several thousand possible comparisons. The resulting p-value of 0.007 shows that the cadmium concentrations in the two data groups are significantly different. This test was run without substituting any numbers such as one-half DL for nondetects, and by expressing all remarked data (such as 0.5J) as between the detection and quantitation limits (0.4 to 0.6). As a nonparametric test, no assumption about the data distribution was required to compute the p-value.

Permutation tests are relatively new in software. Their application to data censored at multiple reporting limits, where remarked data are distinguished from data truly below the reporting limit, is a great step forward. Finally, please note that data measured below the detection limit of 0.4 are represented as a <0.4, or (0 to 0.4), not as less than the quantitation/reporting limit <0.6. Using that higher limit when data are measured below the lower limit is a biased procedure called “insider censoring”, and should be avoided.

Related newsletters you might have missed are on our Newsletter Archive page.
Things People Do With Nondetects that are Just Wrong! April 2011
Interval Censoring - Newer methods for nondetects March 2011
Nondetects in fields outside of environmental science: what it can tell us Dec 2009
and others.

3. Our Top Twelve Tips
Many of you have taken our Applied Environmental Statistics course and so heard of our Top Twelve Tips for environmental statistics. We've brushed them off, polished them up, and are currently listing them on our blog site
http://www.practicalstats.com/news/blog.html
You may also receive a notice when each is posted by following @PracticalStats on Twitter.

'Til next time,

Practical Stats
-- Make sense of your data