Practical Stats Newsletter for March 2011

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

In this newsletter:
1. Registration deadlines for upcoming courses
2. Interval-censored methods for nondetect values
3. Online webinars

1. Registration deadlines for upcoming courses

Registration for two of our three spring courses will increase in price on March 21. As always, online registration is available through the Practical Stats "Upcoming Courses" page (URL below). Registering before March 21 will save your project or employer money. It also allows us to better plan for how many will be attending, have sufficient materials ready, etc., which is the reason for the cost increase. Note that registration for our third course, Applied Environmental Statistics, increases on April 11th,, also not that far away. If you have benefitted from our newsletters or courses, please tell your contacts about them. It makes a difference. Word of mouth is still a major way that people find out about them.

Time Series and Forecasting
for frequently-collected, "real-time" data
April 4-5, 2011 $895 registration before March 21
Homewood Suites Littleton $995 on or after.
Littleton Colorado 80127

Nondetects And Data Analysis
Correctly interpret data below detection limits
April 6-7, 2011 $895 registration before March 21
Homewood Suites Littleton $995 on or after.
Littleton Colorado 80127

Applied Environmental Statistics
Statistics, down to earth
May 2-6, 2011 $1395 registration before April 11
Temple Univ. City Center $1495 on or after.
Philadelphia, PA 19102

You can always find our complete course listing on our "Upcoming Courses" page at http://www.practicalstats.com/new_classes/classes.html

2. Interval-censored methods for nondetect values

Censored nondetect data can be reported and used as either "left-censored" or "interval censored" values. An example of a left-censored format is a "<1". The value is known to be somewhere below 1. Because most environmental data are strictly non-negative, an equivalent interval-censored format can also be used: the value is between 0 and 1, or (0,1). An observation is represented by its range of possible values. Historically, software for nonparametric methods expect censored data to be right-censored "greater-than" values, while software for parametric methods expect data to be input as interval-censored. The textbook *Nondetects And Data Analysis* shows how left-censored nondetect data can be transformed to right-censored values and then used in commercial software offering nonparametric statistical tests. It also shows how to input interval-censored values into parametric tests.

Because nonparametric methods represent data by their order in the data set – ranks, scores or percentiles – a lower boundary for a nondetect such as <1 is not really needed as long as all data possess the same lower boundary. Both a <1 (a "0 to 1") and a <3 (a "0 to 3") can be incorporated into nonparametric survival analysis procedures. But what if lower boundaries are not the same? For example, one observation is a true nondetect (0 to the detection limit) and another an "in-between" value between the detection limit (DL) and reporting limit (RL). For a DL of 1 and a RL of 3, the latter observation is between 1 and 3, not between 0 and 3. These "in-between" values are usually reported as remarked data, something like a "2.5E", where the E indicates that the number is estimated, or has appreciable error. The laboratory knows this observation is above the method detection limit of 1, but has sufficient error that it might be a 2, a 2.1, 2.2, all the way up to a 3, perhaps. How can such information be incorporated into the analysis?

For parametric methods based on maximum likelihood, the answer is simple. Report the data in interval endpoints format. A <1 becomes (0,1) and an in-between value is DL to RL, or (1,3). Maximum likelihood incorporates ranges such as these along with detected values – a detected 10 is a (10,10) – and performs an analysis such as computing regression slopes and intercepts. There is no problem in reporting an observation as the range its value might take.

Software for nonparametric methods might use interval-censored data in one of two ways. First, interval-censored nonparametric procedures are appearing on the software scene. The Turnbull method estimates a median and other percentiles for interval-censored data. If all lower boundaries were 0, the result would be identical to the standard Kaplan-Meier estimate of percentiles. The Turnbull method allows data reported with a different low end such as (1,3) to also be used. Other interval-censored procedures are becoming available, and you should find some of them in the upcoming second edition of *Nondetects And Data Analysis* and in the associated *NADA for R* software package. Unlike R, however, only a few commercial software packages include nonparametric methods for interval-censored data.

The second solution is to rank your data appropriately. For example, suppose the following values were machine readings in the lab for a low-level contaminant:

```
-4  -1  0  0.6  0.8  1.2  1.5  1.8  4   8  13  21
```

and with a DL of 1 and RL of 3, were censored and reported as

```
<1  <1  <1  <1  <1  1.2E 1.5E 1.8E  4   8  13  21
```

These can be ranked in the following way. All five values below the DL are tied with each other and so given the average or median of ranks 1 through 5, or a rank of 3. The three estimated values between the DL and RL are considered tied with each other, but higher than values below the DL. They are given the average of ranks 6 through 8, or 7. All quantified values above 3 are given individual ranks, the same ranks as they would have been assigned if reliable individual values for the lower observations were known. The resulting ranks are

```
 3   3   3   3   3   7    7    7    9  10  11  12
```

which reflect both the ordering of values below and above the DL, and the uncertainty in the exact values of "in-between" data that cannot be quantified.

Reporting data as interval-censored values is perhaps the most straightforward and easy to understand procedure for censored values. It is clear to users, simple to explain on a website where outside persons may obtain shared data, and easy to incorporate into statistical test procedures. It allows laboratory personnel to communicate the uncertainty that "in-between" data possess.

In summary, look for software and methods that allow interval-censored values to be easily incorporated into your method toolbox. Look for training courses such as our *Nondetects And Data Analysis* class coming up next month that shows you how to run these newer interval-censored data procedures.

3. Online webinars

For those of you who haven't been able to make it to one of our courses, we are trying something new. We will offer two webinars this spring on the subject of stats for nondetects, conducted through Midwest Geosciences Group (http://www.midwestgeo.com/). Click on their Webinars link at the left of their home page, or click on the 'Print Webinar Schedule' button.

On April 11th I'll present "Why Subbing One-Half of the Detection Limit is Trouble and What You Can Do Instead". It is aimed at people who substitute some proportion of detection limits for nondetects, and think that this can't hurt too badly. Despite assurances from guidance documents over the years, subbing a value for nondetects can quickly get you in trouble! Recent evaluations have shown that significant errors can result from as few as 5-10% nondetects if they are subjected to substitution methods. Then on May 16 I'll present "Handling Nondetect Data Correctly", an overview of methods now available for performing statistics on left-censored data.

One downside of webinars is that the amount of material that can be presented is limited. These are small bites of the topic, and the above two together might compose about half the material in our two-day *Nondetects And Data Analysis* course. Two upsides of webinars are that

a) they are less expensive, especially when no travel dollars are required, and
b) multiple people can attend at a site for one price, making it even more cost-effective.

Those of you who end up registering for one or both webinars, send an email to us and let us know what you thought of them.

'Til next time,

Practical Stats (Dennis Helsel)
-- Make sense of your data