Practical Stats Newsletter for June 2017

Subscribe and unsubscribe: http://practicalstats.com/news
Archive of past newsletters http://practicalstats.com/news/archive.html
(note new URL)

In this newsletter:

A. Practical Stats Courses
B. Treat Outliers Like Children
C. Newsletter Archive, by topic

A. Practical Stats Courses

Our Applied Environmental Statistics courses are on our online training site:
http://practicalstats.teachable.com/
The two courses separately are each $650 for a 1-year access for one person. Or get both courses together in a bundle for $1200.
More courses, and at least one new free 'podcast' or webinar, are also coming soon.

B. Treat Outliers Like Children

Many environmental scientists have gotten into the habit of performing outlier tests such as Grubbs or Rosner tests, or a more informal process like looking at points that plot individually on a boxplot, to determine whether an observation is an outlier. If it is, they toss the observation away. The problem with this is that in statistics, an outlier is not assumed to be 'bad data'. Outlier tests determine only whether an observation is likely to have been generated from a normal distribution. Similarly, excessive numbers of individual points on a boxplot indicate only that the data set is unlikely to follow a normal distribution. Most field data of water, air, biota and soil chemistry are skewed, and do not look like a normal distribution. There is no reason to suspect that they should. Many disciplines such as hydrology and medical statistics assume that their data are best modeled by skewed distributions such as the Weibull or lognormal. They don't expect to see data following a normal distribution. Neither should we. Labeling observations as 'bad' and removing them based on outlier tests alone is not valid.

One of the Top 12 Tips we have taught in our Applied Environmental Statistics course since 1990 emphasizes this point -- *there is no test for 'badness' in statistics*.

Outliers can have one of three causes:

1. A measurement or recording error;
2. An observation from a different population than most of the data, such as water levels caused by a hurricane rather than routine storms, or a concentration resulting from a brief chemical spill into a river; or
3. The data arise from a single skewed population. This is true for many natural phenomena.

If no error can be detected and corrected, outliers should not be discarded based solely on the data values themselves. A decision to delete observations must come from knowledge of the subject expert that an error has been made, or that data will be examined as two or more separate populations. Rather than eliminating actual (and possibly very important)

data in order to use analysis procedures requiring symmetry or normality, data should be modeled with a skewed distribution fitting the observed values, or outlier-resistant methods should be employed.

Outliers may be the most important points in the data set, and should be investigated further as to their cause. If outliers are deleted, it creates the risk that those who use the data set will only see what they expected to see and may miss gaining important new information.

For example, the plot below is of lead concentrations from 71 observations of Flint Michigan drinking water quality taken in February 2015 (R. Langkjær-Bain, *Significance magazine*, April 2017). This histogram shows that the distribution of concentrations is skewed.

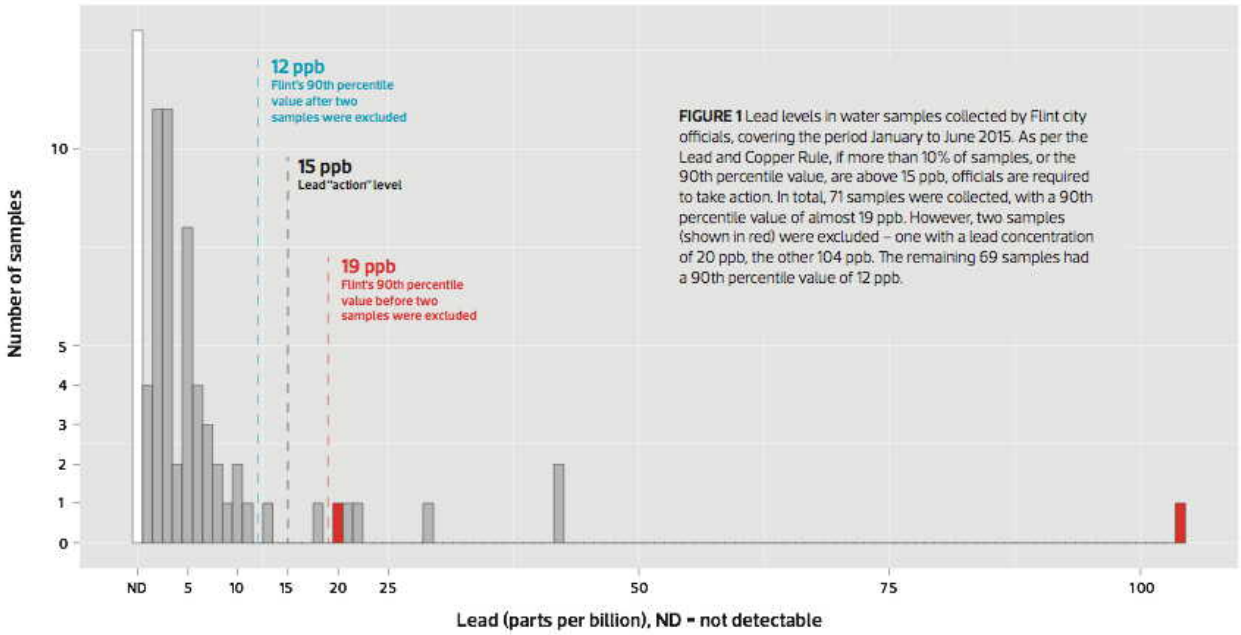

Figure from R. Langkjær-Bain, *Significance magazine*, April 2017

The two red bars represent observations that were thrown out as outliers, taken from locations in Flint where they now know lead was going into solution from the corrosive action of the water supply. Without those two points, the 90th percentile of the remaining observations did not exceed the lead action level, and no additional sampling was triggered, delaying the discovery of lead contamination.

In case you are unsure of how a statistician would have approached this, the article quotes Barry Nussbaum, then president of the American Statistical Association: "There are a lot of statistical methods looking at whether an outlier should be deleted……I don't endorse any of them."

Treat outliers like children. Correct them when necessary, but never throw them out.

---- *Top 12 Tip #2. Practical Stats' Applied Environmental Statistics course*

For all twelve tips from our course, see
http://practicalstats.com/info2use/top12tips.html

C. Newsletter Archive, by topic
With the recent update of our practicalstats.com webpage, the Newsletter Archive now lists our newsletters (from 2003 onward) not just by date, but also by topic. The topics correspond to the training courses we teach. If you haven't recently, browse the archive at its new URL:
http://practicalstats.com/news/archive.html
for interesting information, such as how to compute statistics for data with nondetects. Browsing by topic is far easier than before.

'Til next time,

Practical Stats
-- Make sense of your data