Practical Stats Newsletter for October 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. Statistics for "Small Data", Part 1.
C. Scheduling

A. Upcoming Webinars and Talks

The free webinar "Nondetects And Data Analysis" will be held Nov. 15th as part of the National Water Quality Monitoring Council's webinar series. More information and to sign up:
http://acwi.gov/monitoring/webinars/nondetects_data-analysis_announcement_15nov16.pdf

Our online courses have been in the works for a while, and delayed past what I thought, but they are coming. See section C.

We also offer in-person, onsite training for groups you pull together. See http://practicalstats.com/training/ for details.

B. Statistics for "Small Data" Part 1.

"Big Data", the analysis of hundreds of thousands up to many millions of observations from genomics, internet traffic and other sources, is a popular topic in business and science today. Indeed, Ohio State has established a major in Data Analytics just to prepare "Big Data" analysts. However, trends in environmental monitoring are going in the opposite direction. Fewer and fewer observations are being made in support of critical and costly environmental decisions. This trend is discouraging to say the least, but this month we'll look at two ways to maximize your analysis of small datasets. In December we'll finish by looking at a few additional techniques.

The first tip is to compute exact test statistics with small datasets. Exact tests get their p-values by having computed all the possible outcomes a test could produce for the number of observations taken. The exact p-value is the probability of getting the observed outcome, or an outcome more extreme, when the null hypothesis is true. Tables of exact p-values for nonparametric tests are manually found in (often older) textbooks on nonparametric statistics. Commercial software generally only computes them if the company sells an add-on package for that purpose, at an additional cost. The usual output from commercial statistical software is a "large sample approximation" (LSA) p-value computed by fitting a smooth curve such as the t- or chi-square distribution to the exact distribution of the test statistic. Figure 1 below (from Helsel and Hirsch, 1992) illustrates how a normal distribution is fit to the exact results from a Wilcoxon rank-sum test, here with sufficient data that the curve summarizes the exact results well.

Commercial software will provide p-values from the smooth curve, which for "Small Data" could be off the mark. However, R software, the free statistics package with comprehensive methods, will provide exact test results as its default for many procedures when sample sizes are small – no lookup tables in books required.

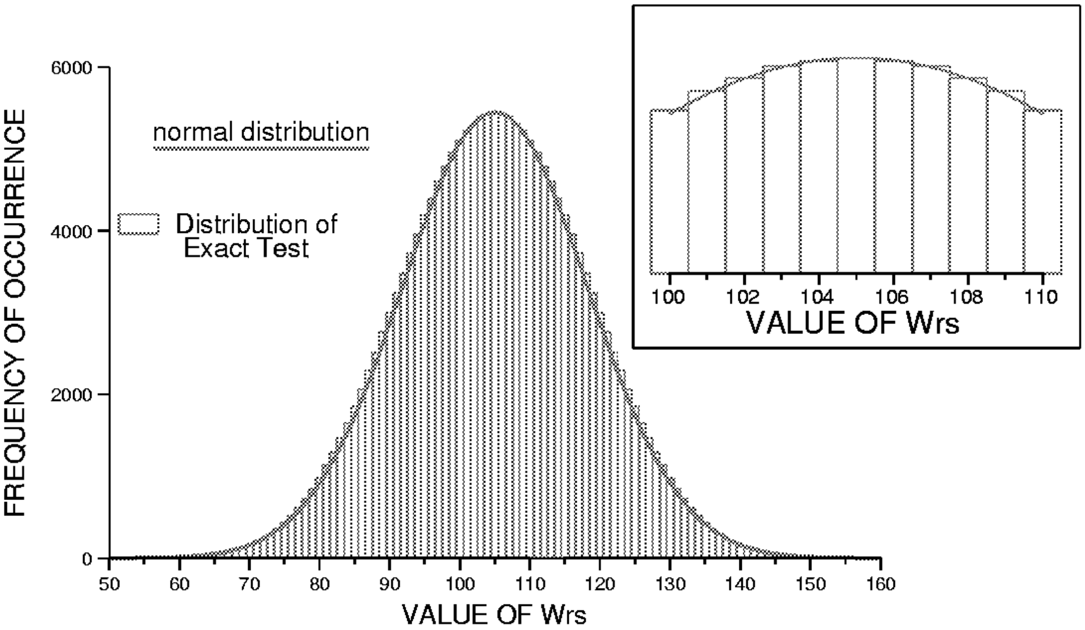

Figure 1. A smooth curve as the large-sample approximation to the p-values for the Wilcoxon rank-sum test.

To show the difference between LSA and exact results, our example dataset measures concentrations in a contaminated shallow aquifer. A remediation effort was implemented to reduce concentrations in waters leaving the site. Data from three upgradient wells above the remediation location and twelve downgradient wells after remediation were obtained. The rank-sum test determines whether downgradient concentrations are significantly lower than upgradient concentrations. With these few samples an exact test would be the best choice, if available. The R commands below first read in the data, perform the exact test, and finally perform a large-sample approximation test.

```
> upgradient <- as.numeric(c(6.900, 3.200, 1.700))
> downgradient <- as.numeric(c(0.390, 0.320, 0.300, 0.305, 0.205,
0.200, 0.195, 0.140, 0.145, 0.090, 0.046, 0.035))
> wilcox.test(downgradient, upgradient, alt="less")

     Wilcoxon rank sum test
data:  downgradient and upgradient
W = 0, p-value = 0.002198
```

The exact test p-value is 0.002, reflecting a strong probability that concentrations decrease in the downgradient group. Had commercial software been used instead, the LSA p-value would instead be about 0.006, produced here in R by specifying to not compute the exact test:

```
> wilcox.test(downgradient, upgradient, alt="less", exact=FALSE,
continuity=TRUE)

     Wilcoxon rank sum test with continuity correction
data:  downgradient and upgradient
W = 0, p-value = 0.00577
```

Commercial software would present the higher p-value while the exact test makes more efficient use of information in the data. In some datasets a smaller exact p-value could be the difference between significance and non-significance.

A second tip for small data is to not decide which test to use, parametric or nonparametric, based on a prior test of normality. Normality tests assume data follow a normal distribution unless proven otherwise. With "Small Data" it is difficult to reject what is only assumed. For the upgradient data:

```
> shapiro.test(upgradient)

     Shapiro-Wilk normality test
data:  upgradient
W = 0.94369, p-value = 0.5425
```

and for the downgradient data:

```
> shapiro.test(downgradient)

     Shapiro-Wilk normality test
data:  downgradient
W = 0.95411, p-value = 0.6977
```

We cannot reject normality for either, yet neither group may follow a normal distribution. Better to use a nonparametric test than use a parametric test on non-normal data, which notoriously has low power to find differences. For example, a t-test did not find a significant decrease in the downgradient group mean.

Reference: Helsel, D.R. and R.M. Hirsch, 1992, Statistical Methods in Water Resources. Elsevier, 522 p.

C. Scheduling

As has been announced here before, the 1992 textbook "Statistical Methods in Water Resources" is being updated and retooled with R for all figures, examples and exercises. The authors (Helsel, Hirsch, Ryberg, Archfield and Gilroy) have been busy, very busy. About half the chapters are out for peer review now, with half going out 'soon'. After that, it all gets pulled together. Bottom line for me (Helsel) is that its meant I've had to put other things on the back burner, including getting our online courses up and running, and even getting out this newsletter in September. If any of you have written 500+ page documents, you understand – a textbook or similar writing project is a big time commitment. All five authors are looking forward to its upcoming release in (reasonable guess: June) 2017.

'Til next time,

Practical Stats
-- Make sense of your data