Practical Stats Newsletter for July 2015

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

In this newsletter:
1. Upcoming 2015 Training
2. Parametric, Nonparametric and Permutation Tests
3. Transitions

1. Upcoming 2015 Training

In-person courses:

Permutation Tests

Aug. 24-25, 2015 $995 through Aug 9, $1095 after.
Altamonte Springs (Orlando area), Florida
Permutation test procedures replace parametric tests like t-tests and ANOVA. Learn about these new, important methods for environmental statistics.
http://practicalstats.com/training/

Applied Environmental Statistics

"Statistics down to earth". A complete survey of statistical methods for environmental data, as well as an introduction to using R statistical software.
Sept. 14-18, 2015. $1495 through August 30, $1595 after.
Lynnwood (Seattle area), Washington
http://practicalstats.com/training/

For Minnesota residents only:

Applied Environmental Statistics
at the Univ. of Minnesota, St. Paul, MN.
Nov. 16-20, 2015
https://drive.google.com/file/d/0B2ukVHiz4qEKdG9KQ2c5SlhjbUU/view?pli=1

2. Parametric, Nonparametric and Permutation Tests

In our previous newsletter (May 2015) we discussed the increased power of permutation tests over their traditional parametric analogs such as the t-test and ANOVA. We presented examples of when the permutation test could see differences between group means, while the parametric test could not, due to non-normality of data. We stated that assuming that things were OK when data did not follow a normal distribution, and running the traditional tests anyway, could often lead to not finding differences between groups that were really there. Permutation tests are a big advance for environmental statistics.

We received a question about how this fits with a paper published by Douglas Johnson, "Statistical Sirens: The Allure of Nonparametrics" (***Ecology, 76(6),*** 1995, **pp.** 1998-2000). Dr. Johnson didn't discuss permutation tests, but contrasted parametric and (rank-

based) nonparametric tests. His comments (see below) concern the assumptions he believes that nonparametric tests require, and the question we received was whether permutation tests make similar assumptions. If they do, how are permutation tests affected by violation of those assumptions, and what are the potential consequences of the violation (e.g., low power, higher risk of Type II error, etc.)?

What a permutation test assumes is a function of what statistic you choose to use. A perm test version of a (nonparametric) rank-sum test, for example, has no distributional assumptions because ranks are used to compute the test statistic. The difference from the usual rank-sum test is that the perm test computes a p-value from thousands of permuted results, rather than the large-sample approximation used by stat software. Good (2012) states that the only assumption for a permutation test, whether ranks or means are used as the target, is that the observations are 'exchangeable' across groups being tested. This is no more than a re-statement of the null hypothesis that the groups' data all come from the same distribution. When running any statistical test you are declaring the null hypothesis is the reasonable statement of 'no signal, no difference', and with a permutation test there are no other distributional assumptions. Yet if the alternate hypothesis is true and the groups differ, how they differ is important whenever testing for differences in means. Consideration of the variation is important, as differences in variation might in some cases obscure the rejection of the null hypothesis that means differ. This is an issue caused by the definition of a mean, and the susceptibility of a mean to outliers, not whether you use a permutation or traditional parametric test.

Perhaps two examples will help. When testing for difference in means of two groups, if the groups have the same variance and their distributions are skewed, a permutation test will correctly determine whether the group means differ without concern for the skewness of the data. This is a big advantage over a t-test. The permutation p-value will be correct even though each group's data do not follow a normal distribution. The t-test's p-value will not. Last month's newsletter presented the equivalent situation for ANOVA.

If the two groups differ in both mean and variance, with the group with higher mean also with higher variance (as is typical for environmental data), permuting the group assignments will produce results representing the null hypothesis that show less difference between group means than if the groups differed only in means, due to the "fog" of the upper group's greater dispersion. When high observations are re-assigned from the upper group to the lower one, they increase the lower group mean a lot because of adding increased variance. The test loses power in comparison to a test on data with identical variance. But you never have the ability to change the variance! Permutation tests handle the difference in variance as well or better than traditional parametric tests when testing for differences in means. It's the consequence of choosing to use a mean as the measure of the group's center.

In summary, permutation tests don't suffer from the loss of power of the normality assumption of t-tests and ANOVA. They may be affected by changes in variance between groups when testing for differences in means, losing power over the situation where the same difference in mean results from groups with identical variance. The loss

of power is no worse, and often better, than the equivalent parametric test in the same situation. If you want to avoid the problem of differing variance entirely, change the scale of the data to ranks and run a nonparametric test. But then you'll be testing a different hypothesis, one of frequency/percentiles rather than mass/totals/mean. Change the scale by taking logarithms and run a permutation test for difference in mean logs. But then you'll be testing a different hypothesis, that the group medians (mean logs transformed back to original units) differ.

Johnson's article makes four main points (plus one or two others off our topic today):
A. Scientists erroneously believe that nonparametric tests require no assumptions about the underlying distribution of data. To test for differences in means, nonparametric tests require an assumption of equal variance of groups, just as parametric tests do.
B. If variances do differ, "the Welch-Satterthwaite version of the *t* test performs well (Wang 1971)".
C. The t-test doesn't require a normal distribution of data, only of their means, and "that property is assured by the Central Limit Theorem, even for relatively small samples, for all but the most perverse data".
D. "By their very nature, nonparametric methods do not specify an easily interpreted parameter…Parameters are generally of most interest, so we should provide estimates of those parameters that are meaningful and applicable to making real decisions." He then goes on to recommend transformations followed by parametric tests if data are skewed, so that the results can still be interpreted using parameters.

My point by point (letter) response is:
First, Johnson doesn't consider permutation tests, so his article isn't applicable to them directly. He is critiquing the use of rank-based nonparametric tests. To answer each of his points:

A. Nonparametric tests are based on frequencies (ranks, percentiles, how often something occurs), so they don't test 'naturally' for differences in means. There's no reason that they should. In order to force them to do so, if you assume that the distribution of both groups is symmetric, then the mean and median are the same, and so a rank-sum test for difference in medians can also be used as test for difference in means. Because the test is based on ranks, an assumption of equal variances is not required for a nonparametric test. But if you're intent is to test differences in means, changing variance will affect any and all tests due to the strong effects of outliers on a mean. It will likely have much less effect in the dimension of ranks, and so for tests such as the rank-sum test, than it had in the original units used by t-tests and ANOVA.

B. The Welch-Satterthwaite adaptation of the t-test for differing variances should always be used. It is the default in statistics software, and far better than the t-test without it, which can lead to incorrect outcomes when variances differ. However, the adaptation costs a lot of power – p-values can be quite high in comparison to a permutation test on the same data. The t-test is often described in the literature as being "robust", which is a technical statistical term. Non-statisticians often believe that "robust" is a general term meaning "really good". However, its technical meaning in statistics addresses only one

specific issue, false positives. The term does not address the t-test's loss of power (false negatives) when data do not follow a normal distribution. For more on this you can read a discussion I co-authored way back in 1988:
https://dl.dropboxusercontent.com/u/39128329/ttestCommentAWRA88.pdf

To simply state that the t-test "performs well" is like saying a 1960 Toyota performs well. Today we would expect fuel injection, seat belts and reduced emissions, not to mention bluetooth and on-board gps. There have been a lot of improvements since the Welch-Satterthwaite's t-test in the 1940s. Today we expect more, including more power and flexibility in a test. A two-group permutation test for means give you many improvements over the classical t-test, even with the Welch-Satthertwaite adaptation.

C. The Central Limit Theorem's determination that the sample mean you've computed follows a normal distribution even though your data do not, is a function of skewness and sample size. Many studies have shown that for the amount of skewness common to environmental data, sample sizes per group must be on the order of 70 observations each for the Central Limit Theorem to allow you to ignore the effects of non-normality on the t-test. USEPA's ProUCL program recommends 100 observations per group before the CLT can be assumed. Strongly skewed data are typical, not "perverse", in environmental science. Because of this, the t-test loses power even for sample sizes not generally considered "small".

D. A median is a parameter. It is different than a mean. Many of the questions we ask of data are frequency questions, such as "does one group exhibit higher values than another?" A nonparametric test answers this question directly. A parametric test on means doesn't actually answer it at all. A median describes the center of data on a frequency scale. Which parameter you test for should be determined by the question you ask. If you transform (by logs, for example) and perform a t-test in the transformed units, this is NOT a test for difference in means in the original units. Often it tests for differences in medians in original units, one reason that it often works quite well. Know what parameters you are actually testing for. Make sure what you test for fits the goals of your study.

Johnson states that a nonparametric rank-sum test actually tests for differences in the cdf, the distribution of groups' data. This is correct, and is one reason to use these tests. For example, suppose your data are <5s for 60% of both groups' data. Above that, one group has mostly <5s in the rest of its data, while the second group has mostly detected values above 5 in its top 40% of data. The rank-sum test can see this type of difference, and the p-value of the test for differences can be significant. The test states that the probability of getting a 'high' value (above 5) is not the same in both groups. The percent of data above 5 in each group might be the appropriate parameter to report, as the median of both groups is <5, and the group means are unknown. The rank-sum test is more general than simply a test for difference in medians, though that is often how it is interpreted. The mean isn't the only parameter available to you, and for situations like this one with nondetects, its not the appropriate parameter. The t-test on data with nondetects cannot be validly computed (especially if you substitute values in order to compute it). Instead

of arguing over 'who's got the best type of test over all of science', use the parameter and test that best matches your objectives. For environmental science, that is more often a percentile (nonparametric test) than a mean (parametric test). Where the important parameter is a mean, use a permutation test rather than the Satterthwaite-Welch t-test to be free from the requirement of a normal distribution.

We will cover these considerations in our Applied Environmental Statistics course this September. You might also take a look at our past newsletters, available in our archive: http://practicalstats.com/news/bydate.html
such as May 2015, August 2014, November 2013, and our 'Urban Legends' discussion of October 2011 to understand related aspects of what this letter has described.

3. Transitions

At Practical Stats our activities fall into four categories:

1. Consulting. Includes data analysis, mentoring, and review/expert witness activities.
2. Webinars. Short (1.5 hour) presentations of statistical topics for environmental studies.
3. Direct classes. Training classes taught directly to companies or agencies. Often including custom content or analysis. Our Minnesota AES class this November is one example, taught only to Minnesota residents through the Univ. of Minnesota.
4. Open classes. Training classes advertised online and open to all, somewhere in the U.S. These are listed in section 1 of this and all our newsletters.

The first three of our activities are profitable. With travel restrictions and decreased training budgets, the open classes have not been profitable for the past several years. We know there is a great need for statistical information and training for environmental scientists, but the opportunity to travel offsite and obtain that training has been severely curtailed. Therefore, our final set of open courses will be taught in the first half (probably first quarter) of 2016. We will continue webinars into 2016 and beyond, and will eagerly come and teach multivariate methods, permutation tests, our week-long AES overview of all of environmental statistics, and our other courses directly at your site. Only the open courses, to which fewer and fewer people have traveled, will be ending. If you have been waiting to take one of our open-registration courses, don't wait any longer. Our Permutation Test class this August and Applied Environmental Statistics course in September will be offered once more in 2016. Our two-day Multivariate class will be offered once more, in 2016. Dates and locations will be on our website by late August, and in our next newsletter in September, so look for them.

After that, invite us over to teach to a group at your site. Or listen to our webinars.

'Til next time,

Practical Stats
-- Make sense of your data