SERIES ON STATISTICS

Nonparametric vs Parametric Tests of Location in Biomedical Research

CHRISTINA M. R. KITCHEN

THE CHOICE OF STATISTICAL TEST HAS A PROFOUND impact on the interpretation of data. Understanding this choice is important for the critical evaluation of the biomedical literature. The question often arises whether to use nonparametric or parametric tests. The *t* test is the most widely used statistical test for comparing the means of 2 independent groups. This parametric test assumes that the data are distributed normally, that samples from different groups are independent, and that the variances between the groups are equal. The most commonly used nonparametric test in this situation is the Wilcoxon rank-sum test (WRST) and the closely related Mann–Whitney *U* test. The WRST assumes that observations from the different groups are random samples (ie, independent and identically distributed) from their respective populations and are mutually independent and that the observations are ordinal or continuous measurements. When there are *k* groups (treatments), the nonparametric test is the Kruskal–Wallis test (KW), a generalization of the WRST. KW is the nonparametric equivalent to analysis of variance (ANOVA). Using nonparametric tests instead of parametric tests brings about 2 questions: 1) What happens if the nonparametric test is used when the parametric assumptions are met?; and 2) What happens when the parametric assumptions are not met?

To answer these questions, one must first discuss the underlying goal of the study. Usually in biomedical applications one is interested in measures of location such as the mean. One can test if the treatment (experimental condition) has an effect (location shift) on the population under study. For example, one may be interested in the effect of treatment(s) on a specific measurement, say cell count, compared to the control. Data of this nature are often analyzed with the *t* test, or if there are $k > 2$ groups, ANOVA. In the parametric case, one tests for differences in the means among the groups. In the nonparametric case, equivalents the location statistic is the median.

The assumptions for the nonparametric test are weaker than those for the parametric test, and it has been stated that when the assumptions are not met, it is better to use the nonparametric test. However, real data are rarely exactly normal.[1–3] Does this mean that one should never use the *t* test? In many datasets seen in the biomedical sciences, there often exist several observations that differ from the others, the so-called outliers. One must also then consider what is the best summary statistic for central tendency. That is, there should be some concept of robustness to assess the properties of the estimators themselves. Robustness, in one sense, refers to the insensitivity of the estimator to outliers or violations in underlying assumptions. One concept of robustness is the breakdown point.[4] The breakdown point is defined as fraction of data that can be arbitrary (corrupted) without making the estimator arbitrarily bad. For example, the sample mean is defined as: $x_1 + x_2 + \ldots + x_n/n$. If we let any one of the observations (say x_n) get arbitrarily large, the mean will become arbitrarily large. This means that even if an investigator has only one large outlier, the mean is arbitrary. Thus, the breakdown point for the mean is 0. The median, which is commonly used when data are skewed or there exist outliers, is defined as the central value in a distribution where above and below lie an equal number of values. Intuitively, one can see that if we let a minority of observations go to infinity, the median will not be arbitrarily bad. The breakdown point of the median is half; this is the highest breakdown point. From the point of robustness and breakdown point, the mean is a good estimator only if the data have zero outliers (no "heavy" tails) and no skewness (symmetry of normal distribution is kept), and there is unimodality. The median is more insensitive to these departures from normality. Nonparametric methods such as the WRST and KW use the median and are thus robust in this sense.

If there exist departures from normality, it seems prudent, in the sense of robustness, to use the nonparametric test. However, one must consider the cost, in terms of power, of applying the nonparametric test when indeed the data are distributed normally and satisfy the other assumptions of the parametric test. With this comes the notion of Asymptotic Relative Efficiency (ARE). The ARE, simply defined, is how many more subjects are needed for the nonparametric test to have equivalent power to the parametric test for a fixed Type I error rate α. If the ARE = 1,

Accepted for publication Jun 30, 2008.

From the Department of Biostatistics, UCLA School of Public Health, Los Angeles, California.

Inquiries to Christina M. R. Kitchen, Department of Biostatistics, UCLA School of Public Health, Los Angeles, CA 90095; e-mail: cr@ucla.edu

0002-9394/09/$36.00
doi:10.1016/j.ajo.2008.06.031

then the 2 tests have equal power for the same number of subjects. AREs <1 indicate that the parametric test is more powerful and AREs >1 indicate that the nonparametric test has more power. The ARE of the WRST vs the *t* test when the underlying assumptions of the *t* test are satisfied is 0.955.[5–7] Similarly, KW vs ANOVA has an ARE of 0.955. However, these nonparametric tests are much more powerful than their parametric counterparts when the underlying distributions are heavy-tailed or have extreme skewness.[5,6,8–10] In some cases the ARE became infinite. Thus, there is minimal power loss associated with the nonparametric tests even when the data are distributed normally, while the power gains of these tests when normality is violated are substantial.

As the sample sizes become infinite, the parametric tests are robust to departures from normality. However, because of cost and potential risks to humans and animals, many of the sample sizes in the biomedical literature are far from infinite. Thus, it is prudent to examine the properties of these estimators when the sample size is small (<25 per group). The small sample properties of the WRST vs the *t* test have been studied extensively.[1,5–9] The WRST has been shown to be as powerful in small samples as the *t* test under the location shift alternatives and can be much more powerful than the *t* test under certain nonnormality conditions.[5,6] Monte Carlo experiments found that for tests of location shift, the WRST was the best test in almost all cases.[8] Further, in some small-sample Monte Carlo simulations the WRST was more powerful than the *t* test even when the two samples were independent, identically normally distributed.[8] The WRST had large power advantages over the *t* test in small sample sizes for distributions that possessed extreme asymmetry or where there existed a point mass at 0.[1] Moreover, under normality conditions with small samples, ANOVA performed only slightly better than KW. However, when the distributions were mixtures of normals, exponential, or double-exponential, KW was substantially more powerful.[10]

Data are often nonnormal in the biomedical sciences[1–3,11] and the sample sizes are often small. In data where there exists skewness, extreme asymmetries, multimodality, or heavy tails, nonparametric tests such as WRST and KW offer a very satisfactory alternative to parametric tests, especially in small samples. Taken together, these results suggest that when the data are distributed normally and all of the other assumptions are met, there is relatively little loss in terms of power to use WRST or KW and there can be almost infinite gains when these assumptions are not met. Because of this, one should consider using the nonparametric test of location for the primary analysis.

THE AUTHOR INDICATES NO FINANCIAL SUPPORT OR FINANCIAL CONFLICT OF INTEREST. THE AUTHOR WAS INVOLVED IN design and conduct of study; data collection; analysis and interpretation of data; and preparation and review of the manuscript.

REFERENCES

1. Bridge P, Sawilowsky S. Increasing physicians' awareness of the impact of statistics on research outcomes: comparative power of the *t* test and Wilcoxon rank-sum test in small samples applied research. J Clin Epidemiol 1999;52:229–235.
2. Hill M, Dixon W. Robustness in real life: a study of clinical laboratory data. Biometrics 1982;38:377–396.
3. Micceri T. The unicorn, the normal curve, and other improbable creatures. Psychol Bull 1989;105:156–166.
4. Donoho D, Huber P, editors. The Notion of Breakdown Point. Belmont, California: Wadsworth, 1983:157–184.
5. Hodges J, Lehman E. The efficiency of some nonparametric competitors of the *t* test. Ann Math Stat 1956;23:169–192.
6. Chernoff H, Savage I. Asymptotic normality and efficiency of certain nonparametric tests. Ann Math Stat 1958;29:927–999.
7. Dixon W. Power under normality of several nonparametric tests. Ann Math Stat 1954;25:610–614.
8. Tanizaki H. Power comparison of nonparametric tests: small sample properties from Monte Carol experiments. J Appl Stat 1997;24:603–632.
9. Neave H, Granger C. A Monte Carlo study comparing various two-sample tests for differences in the mean. Technometrics 1968;10:509–522.
10. Zimmerman D. Increasing the power of the ANOVA F test for outlier-prone distributions by modified ranking methods. J Gen Psychol 1995;122:84–94.
11. Stigler SM. Do robust estimators work with real data. Ann Stat 1977;5:1055–1098.