Worrying about normality

Some recent debates have highlighted the recurring problem of which statistical approach to use in the analysis of non-normally distributed data and/or small unequal data-sets. Received wisdom suggests that,[1,2] when comparing two independent groups in the presence of one or both of the above conditions, the *t*-test may be unreliable and the Wilcoxon-Mann-Whitney (WMW) test is preferable.[3-5] That is, type I and II error rates are affected by violation of underlying test assumptions; type I errors are "liberal", resulting in spurious rejection of the null hypothesis, and power rates are depressed resulting in undetected effects.[6] Modifications of the *t*-test do exist, for both inequality of variance[7] and skewness[8] (similarly for the WMW test)[9] and these have been implemented variously in statistical packages. The important point to note is that such recommendations tend also to be detail non-specific, to the extent that they do not address the question of the degree of non-normality or "how small". That such a conventional strategy of preference for the WMW test for "non-normal data" analysis may lead to strikingly different conclusions was demonstrated by Barber and Thompson using a cost data-set:[3] a p value for the WMW test of 0.011 and for the *t*-test, 0.71. That one would select the p-value according to the underlying hypothesis is unconscionable, but not unknown, in the same manner as the presentation of the standard error (SE) instead of standard deviation (SD) to make the data look "better".[10] Moreover, such a "test dredging" approach, involving multiple standard tests, increases the Type I error beyond the nominal 5% level.[11]

When considering the use of statistical tests such as the t-test and the WMW test, an important and frequently overlooked assumption is that of independence of observations. Assuming independence is highly likely to be reasonable in the clinical trial context, but this is not necessarily so for epidemiological studies, where for example, patients may be clustered by disease state or other criteria. As Cox and Hinkley observed some time ago, "The main emphasis in distribution-free tests is on avoiding assumptions about distributional form. In many applications, however, the most critical assumptions are those of independence." [12]

It is important to realize that the WMW test gives no more protection against false-positive inference than the *t*-test;[4] the WMW test, in its normal approximation, being equivalent to the *t*-test on the ranks of the original variables.[13,14] Moreover, the interpretation of the null hypothesis being tested with the WMW test is not easily described;[2] conventional interpretation (including examples provided by statistical packages[15]) would have it that the null hypothesis is one of equal group medians, but such is not the case; rather it is a test for equality of group mean ranks.[16] Because the WMW test is a test of both location and shape, its interpretation as a test of medians is valid when the only distributional difference is a shift in location.[17] That is, for the MWM test to be "distribution free", under the null hypothesis of equal medians, the two populations being compared are assumed to be continuous and have the same shape.[9] Technically, the above condition will be satisfied if the distribution location parameter is the median (for example a Cauchy distribution); otherwise the location parameter is usually the mean. If the means/medians are similar but the two distributions have different spreads, the WMW test has poor power under the alternative hypothesis. The (Mood) median test also has low power in small samples and is not recommended.[18] That the differences in spread may be as important as (putative) differences in medians is often overlooked. Furthermore, there are different outcomes from WMW test dependent upon the statistical package; these differences relate to the handling of ties, the use of the continuity correction and the use of the asymptotic approximation versus the exact permutation distribution. The latter form of the WMW test would appear to be the preferred option.[15]

To overcome some of these problems, data transformation (log, square root and reciprocal) to achieve approximate normality is often used, but such transformations result in comparisons of geometric (for log transformation[19]) and harmonic (for reciprocal transformation) means and statistical inference in comparing these means cannot be equated with the test of arithmetic means, unless (for geometric means, at least) the variances on the log-scale are equal.[20-22] Due note of the potential loss of power (ranging from 2 to 10%) in analysing transformed data must also be undertaken.[23] Back transformation (to the original scale) may also be problematic for "differences" following square root or reciprocal transformations; with logarithmic transformation, the antilog of a mean log difference gives the ratio of geometric means.[24] Although frequently used in environmental and chemical research, the geometric mean has not been without its critics as an appropriate data summary statistic.[25] Moreover, the geometric mean is a biased estimator of the arithmetic mean and the latter statistic may be appropriate in considering such variables as costs and their surrogates; that is, where consideration of total costs is of importance (total costs = average costs *x* number of patients). The skewness[26] of the distributions of costs and length of stay[27] may mandate, for descriptive purposes, the reporting of summary statistics such as mean and standard deviation, median and range (see for example Esteban *et al*[28]). However, such does not negate the appropriateness of the arithmetic mean for statistical inference.[29]

The above being said, what is known of the performance of the two tests under varying conditions? Lumley *et al*,[2] reviewing a number of studies of *t*-test performance under conditions of non-normality, with sample sizes ranging from as low as 3 to greater than 80, found the performance to be acceptable in terms of Type I and II errors; kurtosis[30] having less impact than severe skewness (the effect of positive skewness actually results in the sampling distribution of the *t* statistic becoming negatively skewed).[31] This also applied to extreme distributional situations of "floor effects" or "discrete mass at zero"; that is, when up to 50% of subjects record zero for the measured variable.[32,33] It also is often forgotten that at sample sizes of 25 to 30 and above, by virtue of the Central Limit Theorem, the sampling distribution of *t* is effectively normal.[34,35] With respect to comparisons with the WMW test, results have been variable, dependent upon the experimental set-up; in particular, the use of "mathematically convenient" distributions or "real life" data sets from different disciplines, including psychology and education.[36] Skovlund and Fenstad,[37] in a simulation study with sample sizes ranging from 5 to 15 and using combinations of equal and unequal variances and sample sizes, and distributions as normal, heavy tailed and skewed, reported that the *t*-test (and the Welch variant, for unequal variances) again had acceptable performance for most of the combinations, except for severely skewed distributions with unequal sample size and unequal variances. The WMW test was shown to be very sensitive to unequal variances (that is, deviations from a pure shift model) and was not recommended for any combination of data characteristics when this condition was present. Under these circumstances, the Welch *t*-test variant and/or data transformation was recommended. Similar results were noted by Zimmerman,[38] who varied both normality and homogeneity of variance; in particular, the Type I error probabilities of the WMW test were more severely distorted with heavy tailed densities, unequal variances and sizes, with larger variances associated with the smaller sample size (n varying from 15 to 40). Bridge and Sawilowsky,[39] investigating the power (ability to detect a false null hypothesis) of the two tests in multimodal, mass at zero and extreme asymmetry distributions, with small n, found a comparative power advantage for the MWM test, which was substantial in some instances; supporting the previous study of Zimmermann and Zumbo.[14] However, these comparisons involved a location shift only and were not subject to multiple violations of assumptions, which, as Zimmerman observed,[38] can produce "anomalous effects not observed in separate violations".

The *t*-test would thus appear to be surprisingly robust to violation of assumptions, and any advantage of the MWM test would appear to be in situations of extreme skewness of underlying distributions (albeit such advantage may be compromised by variance non-homogeneity of the two groups being compared) or extreme outliers. Paradoxically, the WMW test becomes far less robust with increase in sample sizes.[31,35] However, alternatives to these tests are available.[40] On the basis that biomedical research usually involves small samples and proceeds via randomisation of a non-random sample rather than random sampling, and thus the randomisation, not the population model applies, Ludbrook has argued for the use of permutation tests.[41,42] With appropriate software,[43] permutation tests[44] have become a feasible option and noted to have some advantage.[45] A second approach is the non-parametric bootstrap,[46] in which an empirical estimate of the sampling distribution of the statistic in question is obtained by repeated sampling (for example, 1000 times) with replacement from the observed data. A number of recent papers have used bootstrap techniques in the analysis of skewed data.[20,24,47,48] The lack of widespread use of these two alternative approaches to the *t*-test and the WMW test may reflect the previously described lag-time of diffusion of statistical techniques into the medical literature.[49]

The points raised above are illustrated by a consideration of intensive care unit (ICU) cost data in dollars, for two different hospitals, previously reported (Table 1).[50]

Table 1. Intensive care costs for two hospitals

Hospital	*patient number*	*mean*	*median*	*SD*	*min*	*max*
1	410	$5463	$2478	$8767	$242	$69327
2	244	$6366	$3155	$9113	$605	$69426
Total	654	$5800	$2804	$8901	$242	$69426

The costs (both total and for each hospital) demonstrated significant kurtosis (p = 0.001) and skewness (p = 0.001), albeit there was no variance inequality between the hospitals (p = 0.50). Log transformation, in this case, did not effect normality (Shapiro-Wilk test, p = 0.0001) and served only to exacerbate variance disparity (p = 0.006). The difference in mean costs between the hospitals via the *t*-test was non-significant with a p = 0.21, whereas the WMW test suggested a difference between hospital costs at the 0.0001 level. Log transformation of costs resulted in a significant *t*-test (p = 0.0004), but as noted above, such refers to a comparison between geometric means and not a test of the "original" null hypothesis, of equality of arithmetic means, a point reiterated by Zhou *et al*.[51] Using the BCa bootstrap method,[52] no significant difference was noted between mean costs for the hospitals (95% CI of the difference: -$337 to $2552); similarly, the two-sided p value for the (exact) permutation two-sample test, using the raw data as scores, was 0.11. It would appear, therefore, that no difference existed between the (mean) ICU costs of the two hospitals.

We conclude then that formulaic application of statistical tests is inappropriate; careful consideration of the null hypothesis being tested is needed to guide statistical inference.

J. L. MORAN
Intensive Care Unit, Queen Elizabeth Hospital, Woodville, SOUTH AUSTRALIA

P. SOLOMON
University of Adelaide, Adelaide, SOUTH AUSTRALIA

REFERENCES

1. Coyle D. Statistical analysis in pharmacoeconomic studies. A review of current issues and standards. Pharmacoeconomics 1996;9:506-516.
2. Lumley T, Diehr P, Emerson S, Chen L. The importance of the normality assumption in large public health data sets. Annu Rev Public Health 2002;23:151-169.
3. Barber JA, Thompson SG. Open access follow up for inflammatory bowel disease. Would have been better to use t test than Mann-Whitney U test. BMJ 2000;320:1730-1731.
4. Ludbrook J. The Wilcoxon-Mann-Whitney test condemned. Br J Surg 1996;83:136-137.
5. Murray GD. Reply from BJS Statistical Adviser. Br J Surg 1996;83:137.
6. Wilcox RR, Keselman HJ, Kowalchuk RR. Can treatment group equality be improved?: The bootstrap and trimmed means conjecture. Br J Math Stat Psychol 1998;51:123-134.
7. Welch BL. The significance of the differnece between two means when the population variances are unequal. Biometrika 1937;29:350-362.
8. Johnson NJ. Modified t tests and confidence intervals for asymmetrical populations. Journal of the American Statistical Association 1978;73:536-544.
9. Fligner MA, Policello II GE. Robust rank procedures for the Behrens-Fisher problem. Journal of the American Statistical Association 1981;76:162-168.
10. Brown GW. Standard deviation, standard error. Which 'standard' should we use? Am J Dis Child 1982;136:937-941.
11. Gans DJ. The search for significance: different tests on the same data. Journal of Statistical Computing and Simulation 1984;19:1-21.
12. Cox DR, Hinkley DV. Theoretical Statistics. London, Chapman & Hall, 1974.
13. Conover WJ, Iman RL. Rank transformations as a bridge between parametric and nonparametric statistics. The American Statistician 1981;35:124-129.
14. Zimmerman DW, Zumbo BD. Parametric alternatives to the Student t-test under violation of normality and homogeneity of variance. Perceptual & Motor Skills 1992;74:835-844.
15. Bergmann R, Ludbrook J, Spooren WP. Different outcomes of the Wilcoxon-Mann-whitney test from different statistics packages. The American Statistician 2000;54:72-77.
16. Ludbrook J. Statistics in physiology and pharmacology: a slow and erratic learning curve. Clin Exper Pharmacol Physiol 2001;28:488-492.
17. Hart A. Mann-Whitney test is not just a test of medians: differences in spread can be important. BMJ 2001;323:391-393.
18. Freidkin B, Gatswirth JL. Should the median test be retired from general use? The American Statistician 2000;54:161-164.
19. Keene ON. The log transformation is special. Stat Med 1995;14:811-819.
20. Barber JA, Thompson SG. Analysis of cost data in randomized trials: an application of the non-parametric bootstrap. Stat Med 2000;19:3219-3236.
21. Millns H, Woodward M, Bolton-Smith C. Is it necessary to transform nutrient variables prior to statistical analyses? Am J Epidemiol 1995;141:251-262.
22. Zhou XH, Melfi CA, Hui SL. Methods for comparison of cost data. Ann Intern Med 1997;127:1-6.
23. Kingman A, Zion G. Some power considerations when deciding to use transformations. Stat Med 1994;13:769-783.
24. Briggs A, Gray A. The distribution of health care costs and their statistical analysis for economic evaluation. J Health Serv Res Policy 1998;3:233-245.
25. Parkhurst DF. Arithmetic versus geometric means for environmental concentration data. Environmental Science & Technology 1998;92-98.
26. Benjamini Y, Krieger AM. Concepts and measures of skewness with data-analytic implications. The Canadian Journal of Statistics. 1996;24:131-140.
27. Weissman C. Analyzing intensive care unit length of stay data: problems and possible solutions. Crit Care Med 1997;25:1594-1600.
28. Esteban A, Anzueto A, Frutos F, et al. Characteristics and outcomes in adult patients receiving mechanical ventilation: a 28-day international study. JAMA 2002;287:345-355.
29. Barber JA, Thompson SG. Analysis and interpretation of cost data in randomised controlled trials: review of published studies. BMJ 1998;317:1195-1200.
30. DeCarlo LT. On the meaning and use of kurtosis. Psychological Methods 1997;2:292-307.
31. Sutton CD. Computer-intensive methods for tests about the mean of an asymmetrical distribution. Journal of the American Statistical Association 1993;88:802-810.
32. Sullivan LM, D'Agostino RB. Robustness of the t test applied to data distorted from normality by floor effects. J Den Res 1992;71:1938-1943.
33. Sawilowsky SS, Hillman SB. Power of the independent samples t test under a prevalent psychometric measure

distribution. Journal of Consulting & Clinical Psychology 1992;60:240-243.
34. Boneau CA. The effects of violations of assumptions underlying the t test. Psychol Bull 1960;57:49-64.
35. Stonehouse JM, Forrester GJ. Robustness of the t and U test under combined assumption violations. Applied Statistics 1998;25:63-73.
36. Sawilowsky SS, Blair RC. A more realistic look at the robustness and type II error properties of the t test to departure from population normality. Psychol Bull 1992;111:352-360.
37. Skovlund E, Fenstad GU. Should we always choose a nonparametric test when comparing two apparently nonnormal distributions? J Clin Epidemiol 2001;54:86-92.
38. Zimmerman DW. Invalidation of parametric ans nonparametric statistical test by concurrent violation of two assumptions. The Journal of Experimental Education 1998;67:55-68.
39. Bridge PD, Sawilowsky SS. Increasing physicians' awareness of the impact of statistics on research outcomes: comparative power of the t-test and and Wilcoxon Rank-Sum test in small samples applied research. J Clin Epidemiol 1999;52:229-235.
40. Zhou XH, Li C, Gao S, Tierney WM. Methods for testing equality of means of health care costs in a paired design study. Stat Med 2001;20:1703-1720.
41. Ludbrook J. Advantages of permutation (randomization) tests in clinical and experimental pharmacology and physiology. Clin Exper Pharmacol Physiol 1994;21:673-686.
42. Ludbrook J, Dudley H. Why permutation tests are superior to t and F tests in biomedical reearch. The American Statistician 1998;52:127-132.
43. StatXact 4.01 for Windows: Statistical software for exact nonparametric inference. 1999 Cytel Software Corporation Cambridge MA
44. Good.P. Permutation tests: a practical guide to resampling methods for testing hypotheses. Second ed. New York: Springer-Verlag; 2000.
45. Cohen ME, Arthur JS. Randomization analysis of dental data characterized by skew and variance heterogeneity. Community Dent Oral Epidemiol 1991;19:185-189.
46. Efron, B and Tibshirani, R. An introduction to the bootstrap. New York: Chapman and Hall; 1993.
47. Desgagne A, Castilloux AM, Angers JF, LeLorier J. The use of the bootstrap statistical method for the pharmacoeconomic cost analysis of skewed data. Pharmacoeconomics 1998;13:487-497.
48. Rascati KL, Smith MJ, Neilands T. Dealing with skewed data: an example using asthma-related costs of medicaid clients. Clin Ther 2001;23:481-498.
49. Altman DG, Goodman SN. Transfer of technology from statistical journals to the biomedical literature. Past trends and future predictions. JAMA 1994;272:129-132.
50. Moran JL, Peisach AR, Solomon P. Modelling total costs in adult intensive care units: new models for old questions. Intensive Care Med 2001;27:S142.
51. Zhou XH, Gao S, Hui SL. Methods for comparing the means of two independent log-normal samples. Biometrics 1997;53:1129-1135.
52. Carpenter J, Bithell J. Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. Stat Med 2000;19:1141-1164.