Biotechnology and Molecular Biology Reviews Vol. 4 (6), pp. 128-131, December, 2009
Available online at http://www.academicjournals.org/bmbr
ISSN 1538-2273 © 2009 Academic Journals

Standard Review

Statistical analysis of the application of Wilcoxon and Mann-Whitney U test in medical research studies

U. M. Okeh

Department of Industrial Mathematics and Applied Statistics, Ebonyi State University, Abakaliki Nigeria. E-mail: umokeh1@yahoo.com. Tel: +2348034304278.

Accepted 11 December, 2009

Although non-normal data are widespread in biomedical research, parametric tests unnecessarily, predominate in statistical analyses. Five biomedical journals were surveyed and for all studies which contain at least the unpaired t-test or the non-parametric Wilcoxon and Mann-Whitney U test - investigated the relationship between the choice of a statistical test and other variables such as type of journal, sample size, randomization, sponsoring etc. The non-parametric Wilcoxon and Mann-Whitney U were used in 30% of the studies. In a multivariable logistic regression the type of journal, the test object, the scale of measurement and the statistical software were significant. The non-parametric test was more common in case of non-continuous data, in high-impact journals, in studies in humans, and when the statistical software is specified, in particular when SPSS was used.

Key words: Wilcoxon and Mann-Whitney U test, univariate analyses, non-parametric test, logistic regression.

INTRODUCTION

When looking into the medical literature one gets the impression that parametric statistical methods such as Student's t-test are common standard, although the underlying normal assumption is often not tenable, especially for small or moderate sample sizes. On the one hand, empirical work has shown that deviations from a normal distribution are frequent even for continuous data (Micceri, 1989). According to Nanna and Sawilowsky (1998), normality is the exception rather than the norm in applied research. However, for large sample sizes one may rely on the central limit theorem and apply a test designed for normally distributed data. On the other hand, ordinal data are widespread in biomedical research (Rabbee et al., 2003). For such data non-parametric tests based on ranks are appropriate, but the statistical analysis is often not performed properly, as shown e.g. by Jakobsson (2004) for the analysis of ordinal data in nursing research. Sometimes a transformation is applied in order to normalize continuous, but non-normal data. However, in case of non-normal data it is preferable to perform a nonparametric test. Transformations can often not be applied since the transformation "must be motivated from previous experimental or scientific evidence. Unless determined a priori, transforms can be misused to inflate or mitigate observed significance in a spurious fashion" (Piegorsch and Bailer, 1997 p. 130). Furthermore, the hypotheses before and after the transformation may differ (Games, 1984). Hence, the use of transformations for the sole purpose of complying with the assumptions of parametric tests is dangerous (Wilson, 2007). Investigation was made on how frequent the t-test and its nonparametric competitor, the Wilcoxon and Mann-Whitney (WMW) U test, are used in medical research. It is enquired which factors and variables are important for the choice between the non-parametric WMW test and the parametric t-test for research that compare two independent groups, published in medical journals with different scopes and impact. It will be discussed whether the decision for one of the methods is appropriate or not.

METHODS

All original work related to medical studies published in 2004 in five biomedical journals was surveyed. The three journals American Journal of Physiology (Heart Circ. Physiol.), Annals of Surgery, and Circulation Research were considered because they were also included in a previous study (Ludbrook and Dudley, 1998). In addition, The Lancet and The New England Journal of Medicine were included in my study. These journals were categorized into two groups with different topics and impact factors (Table 1). Each paper was thoroughly checked, on whether it included original material on not yet published data, irrespective of medical subject, study design or size/format of the paper. For the analyses presented here all research, which contain at least the unpaired t-

Table 1. Included journals and number of studies.

Journal	Surveyed studies	Included studies	Impact factor (2004)	Type of journal
American Journal of Physiology (Heart Circ. Physiology.)	645	251(38.9%)	3.5	Primarily specialized in subject
Annals of Surgery	227	83(36.6%)	5.9	,,
Circulation Research	343	143(41.7%)	10.0	,,
The Lancet	391		21.7	Articles of diverse topics
The New England Journal of Medicine			38.6	,,

test or the WMW test, were included. In addition to the test statistic the following factors and variables were also inspected: type of journal, sample size, kind of test objects, scale of measurements, information about randomization, sponsoring by pharmaceutical companies, and the used statistical software. Analyses were performed with logistic regressions. When the software used for analysis cannot perform both the t-test and the WMW test the respective study was excluded from the logistic regression analysis. The total sample size was categorized into three categories with an approx. equal number of research studies (<15, 15 - <50, ≥50). Odds ratios (OR) and their 95% confidence intervals (95% -CI) were estimated by logistic regressions. A p-value ≤0.05 was considered as significant. Because of the exploratory nature of my research no multiplicity adjustment was applied (Neuhäuser, 2006).

RESULTS

In total, 1879 publications were surveyed, and 630 research studies could be included in the analyses (Table 1). Altogether the use of the unpaired t-test predominates in studies where two groups were compared. In 112 studies (18%) only the WMW and in 444 studies (70%) only the unpaired t-test is used; 74 times (12%) both tests are applied within one study. Please note that the two tests may be used to analyse different variables, however, it was also found that identical variables were analysed with both tests. In the logistic regressions presented below the studies without the WMW test are compared with the research studies with the WMW test. Two of the 630 studies were excluded from the logistic regression analyses because the specified software cannot perform the WMW test. The univariate analyses show significant relationships between the use of the WMW test and the journal type. The WMW test is more common in the diverse and high-impact journals The New England Journal of Medicine and The Lancet ($p \leq 0.001$, OR=5.21, 95% -CI: 3.53 - 7.69). Moreover, the WMW test is more common in studies in humans ($p \leq 0.001$, OR = 6.44, 95% -CI: 4.42 - 9.38), and, not surprisingly, in research studies with non-continuous variables ($p \leq 0.001$, OR = 8.49, 95% -CI: 4.73 - 15.27). In addition, the statistical software used is significantly related to the choice between the two statistical tests ($p \leq 0.001$). In particular, the WMW test is more common when one of the two common software packages SPSS ($p = 0.004$, OR = 4.64, 95% -CI: 2.48 - 8.69) and SAS ($p = 0.030$, OR = 4.34, 95% -CI 1.96 - 9.61) is used. Another significant relationship was found regarding information about randomization ($p \leq 0.001$, Odds Ratio (OR) = 2.44, 95% -Confidence Interval (CI): 1.70 - 3.50). The WMW test seems to be more common when the study is sponsored by a pharmaceutical company (p=0.028, OR=2.32, 95% -CI: 1.10 - 4.90). The sample size was also significant in the univariate logistic regression ($p \leq 0.001$). In particular, the WMW test was applied more often in case of large samples (that is, $n \geq 50$) than in case of small samples (that is $n < 15$) ($p = 0.001$, OR = 5.88, 95% -CI: 3.68 - 9.39). Obviously, the different factors are not independent. Therefore, a multivariable logistic regression was applied in order to confirm the univariate results. The type of journal, the test object (research studies in humans or in other subjects), the scale of measurement (continuous or not) and the statistical software used remained significant (Table 2). The factors randomization, sponsoring and the categorized sample size are no longer significant. With regard to the software, SAS is no longer significant, either. The multivariate regression gives a significantly larger probability for performing the WMW test for SPSS, only. Sometimes, to be precise, in 57 studies, a reason is specified for using the WMW test. The most common reasons are “non-normal data” and “categorical data”. Further correct reasons are “requirements for t-test not fulfilled” and “small sample sizes”. However, the latter reason is correct only when applying the exact (permutation) version of the WMW test. There are also reasons that are problematic from a statistical point of view: In four research studies the WMW test was applied before or after the t-test, at least partly because the t-test was not significant. In one further study the WMW test was used because an observed heterogeneity in variances. However, the WMW test cannot guarantee the significance level in case of unequal variances (Kasuya, 2001). Moreover, the specified reason “in order to compare medians” is correct only if a pure location shift between the two distributions can be assumed. As mentioned above, one may rely on the central limit theorem when sample sizes are large and, consequently, one may apply a parametric test such as the t-test. However, in 395 out of the considered 630 research stu-

Table 2. Results of the univariate and multivariable logistic regressions.

Factor	Reference category	Univariate analysis				Multivariate analysis (n = 590)		
		n	p-value	OR	95%-CI	p-value	OR	95%-CI
Total sample size	<15	594	≤0.001			0.731		
15-<50			0.058	1.63	0.98-2.71	0.705	1.03	0.59-1.80
≤50			≤0.001	5.88	3.68-9.39	0.429	1.27	0.66-2.45
Randomization	No random or not specified	628	≤0.001	2.44	1.70-3.50	0.313	1.25	0.81-1.95
Sponsoring	No sponsoring[1]	628	0.028	2.32	1.10-4.90	0.631	0.80	0.33-1.98
Type of subject	Other than humans	624	≤0.001	6.44	4.42-9.38	0.012	2.08	1.18-3.67
Software used	Not specified	628	≤0.001			0.008		
SAS			0.030	4.34	1.96-9.61	0.528	2.08	0.86-5.06
SPSS			0.004	4.64	2.48-8.69	0.027	3.18	1.55-6.53
Other			0.007	1.19	0.70-2.04	0.175	1.18	0.65-2.17
Scale of measurement	Not only continuous variable	628	≤0.001	0.12	0.07-0.21	≤0.001	0.26	0.13-0.50
Journal Type	Primary specialized	628	≤0.001	5.21	3.53-7.69	0.004	2.25	1.30-3.87

[1]No sponsoring by a pharmaceutical company.

Table 3. Frequencies of study subject by scale of measurement.

		Study subject	
		Human	Other test objects
Scale of measurements	Only continuous variables	162	401
	Not only continuous variables	60	3

dies the (total) sample size is less than 50. In 89% (353) of these research studies with low sample size the t-test was applied, sometimes in addition to the WMW test (34 studies). In the remaining 319 studies with low sample size the t-test, but not the WMW test, was used. However, in 317 out of these 319 studies (99%) there are continuous variables. Hence, given the relatively high robustness of the t-test to skew continuous distributions (Posten, 1978), the basic assumptions seem to be fulfilled in the vast majority of studies when applying the t-test. In case of more than two groups the Kruskal-Wallis test can be applied as a non-parametric test instead of the WMW test. When considering the 1879 surveyed publications the Kruskal-Wallis test was applied in 53 research studies. Many of these studies have a low sample size smaller than 50 (23 studies) and/or non-continuous data (18 studies). The parametric analogue, an analysis of variance (ANOVA), was found in 658 studies. However, these 658 studies cannot be compared with the 53 studies with a Kruskal-Wallis test because an ANOVA is much more flexible than the Kruskal-Wallis test and can also be applied in studies with more complex designs.

DISCUSSION

The assertions some authors made about their decisions for the WMW and the attributes of the published data indicate that the scale of measurement is the primary factor for a decision in favour of a non-parametric test. However, there are three further factors that remained significant in the multivariable logistic regression.

The study subject is one of these significant factors. The WMW test is more often used in studies in humans. However, in these studies non-continuous variables are more common as well (Table 3). Furthermore, the software has a significant influence. A further significant factor is the type of journal. A possible explanation is that the high-impact journals have a more detailed statistical review and that they may reject a paper because of an inappropriate statistical analysis. In line with this, studies published in journals with high impact factors often contain a more detailed methodical description compared to studies published in other journals. Please note in this context that The New England Journal of Medicine says in its instructions for authors that “nonparametric methods should be used to compare groups when the distribution of the dependent variable is not normal” (http://authors.nejm.org/help/newms.asp). In addition to The Lancet and The New England Journal of Medicine we included the three journals American Journal of Physiology (Heart Circ. Physiol.), Annals of Surgery, and Circulation Research in our study. These three latter journals were also included in a previous study (Ludbrook J, Dudley H (1998). This sample of five journals is not

necessarily representative for the multitude of biomedical journals. However, we are able to compare our results towards the work of Ludbrook and Dudley (1998). This comparison indicates that the behaviour of medical scientists with parametric and non-parametric tests did not change considerably. Ludbrook and Dudley's [8] findings about the handling with statistical methods can be approved even ten years later. Given the higher efficiency of non-parametric tests for non-normal data (Lehmann, 1975), non-parametric tests such as the WMW test should be applied more often, especially when the sample size is not very large. In other areas of life sciences the WMW test seems to be more common. Ruxton (2006) surveyed one volume of the journal Behavioral Ecology. The WMW test was applied in 21/33 = 64% of the papers that used the two-sample t-test and/or the WMW test.

REFERENCES

Games PA (1984). Data transformation, power, and skew: a rebuttal to Levine and Dunlap. Psychol. Bull. 95: 345-7.

Jakobsson U (2004). Statistical presentation and analysis of ordinal data in nursing research. Scand. J. Caring. Sci.18(4): 437- 40.

Kasuya E (2001). Mann-Whitney U test when variances are unequal. Anim. Behav. 61(6): 1247-1249.

Lehmann EL (1975). Non-parametrics: Statistical methods based on ranks. San Francisco, CA: Holden-Day.

Ludbrook J, Dudley H (1998). Why permutation tests are superior to T and F tests in biomedical research. Am. Stat. 52(2): 127-32.

Micceri T (1989). The unicorn, the normal curve, and other improbable creatures. Psychol. Bull.105: 156-66.

Nanna MJ, Sawilowsky SS (1998). Analysis of Likert scale data in disability and medical rehabilitation research. Psychol. Methods. 3: 55-67.

Neuhäuser M (2006). How to deal with multiple endpoints in clinical trials. Fundam Clin. Pharmacol. 20(6): 515-23.

Piegorsch WW, Bailer AJ (1997). Statistics for environmental biology and toxicology. London, England: Chapman and Hall.

Posten HO (1978). The robustness of the two-sample t-test over the Pearson system. J. Stat .Comput. Simul. 6: 295-311.

Rabbee N, Coull BA, Mehta C, Patel N, Senchaudhuri P (2003). Power and sample size for ordered categorical data. Stat. Methods Med. Res. 12(1): 73-84.

Ruxton GD (2006). The unequal variance t-test is an underused alternative to Student's t-test and the Mann-Whitney U test. Behav. Ecol. 17(4): 688-90.

Wilson JB (2007). Priorities in statistics, the sensitive feet of elephants and don't transform data. Folia Geobot. 42: 161-7.