Journal of Communication Disorders 58 (2015) 158–168

Contents lists available at ScienceDirect

Journal of Communication Disorders

The *t* test and beyond: Recommendations for testing the central tendencies of two independent samples in research on speech, language and hearing pathology

Toni Rietveld *, Roeland van Hout

Centre of Language Studies, Radboud University Nijmegen, Erasmusplein 1, 6525 HT Nijmegen, The Netherlands

ARTICLE INFO

Article history:
Received 4 November 2014
Received in revised form 9 June 2015
Accepted 7 August 2015
Available online 20 August 2015

Keywords:
Statistical analysis
Two independent samples

ABSTRACT

Purpose: In this Tutorial we compare current practice of the analysis of data obtained in designs involving two independent samples with new developments in statistics and evidence on the behavior of conventional statistics. We included *t* tests, non-parametric alternatives, such as the *Wilcoxon–Mann–Whitney* test, and recently developed approaches, known as *bootstrapping* and *randomization* tests. The relative use of the different statistics is illustrated on the basis of counts carried out in three journals on disordered communication in the time interval 2005–2013: *Clinical Linguistics & Phonetics, Journal of Communication Disorders* and *Journal of Speech, Language and Hearing Research.* A number of recommendations are given to guide the researcher in the presentation and analysis of her/his data.

Conclusions: The main messages are (a) that researchers should present more relevant features of their data (means, medians, SD, skewness, tailedness, outliers etc.), (b) not routinely use conventional non-parametric tests like *Wilcoxon–Mann–Whitney* test in case one or more of the assumptions of *t* tests are not met, and (c) should consider using less conventional, but robust statistics which have been developed and tested in the last decades.

1. Introduction

In his classical article "The unicorn, the normal curve, and other improbable creatures" Micceri (1989) claimed that normally distributed populations hardly ever occur in the behavioral sciences; to our knowledge this claim has not been contested thus far. As a large part of speech–language–hearing research (SLHR) can be seen as member of the family of behavorial sciences, we may assume that Micceri's claim is also valid for language and speech pathology, as the claim was found to be valid for very different subdisciplines in medical research as well, see Bridge and Sawilowsky (1999), Grissom (2000) and Fagerland and Sandvik (2009b). We even dare to assume that the characteristics of the target populations and samples of SLHR deviate more from what is often seen as ideal from the perspective of statistical analysis than in 'average' behavioral research. Skewness of the distributions is not exceptional, tails are often heavy, variances not equal and sample size(s) relatively small especially of those obtained in minority or small languages and infrequent disorders. Summarizing, the situation for research in SLH does not seem very favorable for adequate statistical analysis, including data obtained in

* Corresponding author. Tel.: +0031 24 3612900; fax: +0031 24 3612907.
E-mail address: A.Rietveld@let.ru.nl (T. Rietveld).

http://dx.doi.org/10.1016/j.jcomdis.2015.08.002
0021-9924/

two-sample designs. We have to underline the word 'seem', as hardly ever researchers explicitly mention the characteristics of their populations in the terms mentioned above, let alone publish tests of them (only mentioning SDs is not enough). When two sets of non-count data are obtained in a design with two independent samples many researchers use a *t* test. Often, however, a *nonparametric* alternative is chosen, like the well-known *Wilcoxon–Mann–Whitney* test (*WMW*; sometimes called *Wilcoxon W* or *Mann–Whitney U test*) for cases where one or more of the assumptions of the *t* test are not met, like normality of the distributions of the populations and equality of their variances. The term 'non-parametric' suggests that no assumptions have to be made on the characteristics (parameters) of the distributions involved from which the samples are taken (Siegel & Castellan, 1988); this is not correct as shown by Hayes (2000). For the *WMW* test the H0 is in fact the identity of the two distributions from which the samples are drawn.

In our contribution we will focus on the parametric *t* test and its variants. In order to assess current statistical practice in the domain of speech and language pathology we counted the frequencies of use of parametric and non-parametric statistics for designs with two independent in three journals, covering the years 2005 to 2013. The overview for the whole period is given in Table 1. We restricted our counts to straightforward examples of comparisons of two measures of central tendency, excluding post-hoc comparisons, *t* tests associated with other procedures and analyses of variance with just one factor and two levels.

Two criteria were used to choose these journals: (a) They should belong to the first 15 of a ranking list of Journals in Audiology, Speech & Language Pathology, based on a 5-year Impact factor, with ranks evenly distributed over this list, and (b) The subjects covered by these journals should be general rather than specific. Thus the three journals used in our review can be regarded as quite representative for the field of speech–language–hearing research.

The search terms used were: (1) two independent samples design, (2) *t* test, Welch *t* test, (3) *WMW* test (and alternative names used for this test) and (4) names of the other tests listed in Appendix A. These other tests were hardly ever or never used, and are consequently not mentioned in Table 1.

On the whole the percentages of use of parametric tests outnumber those of non-parametric tests. Arguments for the use of non-parametric tests were not often given. We frequently came across arguments like: "small number of observations", "non-normality of the distribution(s) of data", "unequal sample sizes", "unequal variances", without further details, tests or references to the statistical literature. We will see that robust (i.e. tests which are not sensitive to violations of assumptions) and well-established parametric alternatives are often available. These tests are listed in Appendix A. We hardly ever saw an article in which one of the recently developed robust tests for two samples had been used.

Statistical simulation studies have made available much more information on the pros and cons of different statistical tests, as a function of quite a large number of aspects, such as measurement level (ordinal vs. interval data), normality and symmetry of distributions (skewness), heavy tails, kurtosis, equality of variances, (equal) sample size(s), etc. Primary outcome variables of these simulation studies are the extent to which nominal alpha levels are maintained (probability of a Type I error), the extent to which power is achieved (1—probability of a Type II error) and the relative efficiency of the test (=the sample size needed for test B to achieve the same power as test A). One of the first publications in which simulations were used to assess the behavior of *t* tests in conditions in which basic assumptions were not met was that of Boneau (1960). He showed that violation of the assumption of normality did not seriously affect the probability of Type I errors in *t* tests for independent samples. This publication was followed by a large number of articles, the general tendencies of which we summarize in the following sections.

Next to investigations into the behavior of tests in terms of Type I and Type II errors, other simulation studies focused on the value of preliminary tests to guide researchers toward the use of the right statistical test. Well-known preliminary tests are, for instance, *Levene's* test for equality of variances (used in a variety of between-subject designs in which the assumption of equal variances is important for statistics like the *t* test) and the *Shapiro–Wilk* test for normality. These tests turn out to be not as powerful as is often assumed. This state of affairs made Ruxton (2006) remark that "it is generally unwise to decide whether to perform one statistical test on the basis of the outcome of another".

The choice of a test is also directly related to the null hypothesis to be tested. For the conventional *t* test it is simply the equality of the means of the continuous distributions from which the samples are drawn (their "central locations"). We will see that adherence to this hypothesis is not always favorable to sound testing in terms of Type I and Type II errors; therefore it is recommended to use other hypotheses in adverse conditions like small samples from skewed distributions (Neuhäuser &

Table 1
Numbers of articles in which the parametric test and/or its non-parametric alternatives were applied on data obtained in two independent samples designs, in three journals over the period 2005–2013. Tests associated with other procedures (like post-hoc comparisons in ANOVAs) are not reported.

Independent samples					
Journal	Count and % within journal	*t* test	Welch/Satterthwaite *t*	Wilcoxon–Mann–Whitney (*WMW*)	Total
Clinical Linguistics & Phonetics	Count	110	4	97	211
	%	57%	2%	41%	
Journal of Communication Disorders	Count	120	1	40	161
	%	75%	1%	24%	
Journal of Speech, Language and Hearing Research	Count	355	9	110	474
	%	75%	2%	23%	