Statistics and Probability Letters 109 (2016) 118–123

Contents lists available at ScienceDirect

Statistics and Probability Letters

journal homepage: www.elsevier.com/locate/stapro

Characterising transitive two-sample tests

Thomas Lumley [a], Daniel L. Gillen [b,*]

[a] Department of Statistics, Science Building 303, University of Auckland, Auckland, NZ, USA
[b] Department of Statistics, 2226 Donald Bren Hall, University of California, Irvine, CA 92697-1250, USA

ARTICLE INFO

Article history:
Received 1 November 2015
Received in revised form 9 November 2015
Accepted 9 November 2015
Available online 17 November 2015

Keywords:
Preference relation
Preorder
Transitivity
Utility representation

ABSTRACT

Implicit in two-sample testing is the idea that a test consistently orders possible distributions. We characterise tests that are consistent with some ordering on distributions, showing all are tests for equality of some univariate real-valued functional in the two samples.

1. Introduction

In testing scientific hypotheses we are typically interested in more than equality or inequality of distributions. A common inferential procedure is to perform (double) one-sided tests to compare two empirical distributions with the end result being a statement that one is "better" in some way: values tend to be larger, more dispersed, more peaked, have more modes, et cetera. That is, our scientific objective is not simply to reject the null hypothesis, but to say which direction it is rejected in. Implicit in this practice is the idea that a two-sample test orders possible distributions in some consistent fashion.

In this paper we consider the general setting where an inferential procedure is based on a test statistic U that can be written as a functional $U = U(\mathbb{F}, \mathbb{G})$ where $\mathbb{F}$, $\mathbb{G}$ range at least over all pairs of distributions with finite support. If this test statistic is used to make pairwise comparisons among three distributions, there could be as many as $2^{\binom{3}{2}} = 8$ possible sets of results, but only $3! = 6$ possible orderings. Thus some sets of test results would then be inconsistent with any ordering of the distributions. The potential for paradox increases with the number of distributions, and also increases if distributions are allowed to be equivalent with respect to the ordering.

Many familiar tests are guaranteed to yield results that are consistent with a proper ordering of distributions. Most notably, if the test is based on the difference in some one-dimensional summary of the distributions, the results of pairwise comparisons must be consistent with the ordering of the distribution by this summary. For example, the Welch version of the two-sample t-test (Welch, 1947) is a test for difference in means. The test statistic is zero if the means in two samples are the same, positive if the first mean is higher, and negative if the first mean is lower. If the test statistic is positive comparing $\mathbb{F}$ and $\mathbb{G}$, and positive comparing $\mathbb{G}$ and $\mathbb{H}$, it must be positive comparing $\mathbb{F}$ and $\mathbb{H}$. The Welch t-test is also consistent: in large samples the test has level α if the two means are the same and has power increasing to 1 if the means are not the

* Corresponding author. Tel.: +1 949 824 9862; fax: +1 949 824 9863.
E-mail address: dgillen@uci.edu (D.L. Gillen).

http://dx.doi.org/10.1016/j.spl.2015.11.005
0167-7152/

same. The original (Student) t-test is also test of difference in means, but within families of distributions that have the same variance. Like the Welch t-test, it will give a self-consistent ordering of any set of distributions. In large samples it has power increasing to 1 if two distributions have different means, but has level α only when variances are the same.

The Wilcoxon test, on the other hand, need not give a self-consistent ordering of a set of distributions. The test is routinely used to decide that one distribution tends to be larger than another, but it is possible given three samples X, Y, and Z to conclude that X is larger than Y, Y is larger than Z and Z is larger than X. That is, the Wilcoxon test is not *transitive*. Indeed, the Mann–Whitney U-statistic version of the test is based on the bivariate functional $P(X < Y)$ and non-transitivity paradoxes for this statistic were described by Condorcet in his 1785 *Essay on the Application of Analysis to the Probability of Majority Decisions* (Berg, 2006). The relationship between voting paradoxes and orderings on probability distributions appears to have been first pointed out by Efron, as described in Martin Gardner's *Mathematical Games* column (Gardner, 1970). More recent statistical discussions include Thangavelu and Brunner (2006) and Brown and Hettmansperger (2002).

Although non-transitivity of the Wilcoxon–Mann–Whitney (WMW) test is long-established, most statisticians (>95% in informal sampling by the authors over several years) are unaware of this fact. There is little explicit mention in the statistical literature and essentially no mention in textbooks of the fact that the WMW test does not order distributions in the same way as any one-dimensional summary statistic.

The lack of transitivity when ordering by $P(X < Y)$ has also been discussed in the context of the Pitman closeness criterion, where X and Y are deviations of an estimator from the underlying parameter. An estimator $\hat{\theta}_1$ is Pitman closer (to a parameter θ) than a competing estimator, $\hat{\theta}_2$, if

$$P\left(\left\|\hat{\theta}_1 - \theta\right\| < \left\|\hat{\theta}_2 - \theta\right\|\right) > \frac{1}{2}.$$

In this context non-transitivity is almost universally regarded as a definite disadvantage of Pitman closeness, although opinion is divided as to whether it is a fatal disadvantage, as shown by the discussants of Robert et al. (1993).

Transitivity is important in normative theories of rationality. For example the first of Savage's axioms (Savage, 1954) is that everyone has some linear weak order giving preferences for distributions, and even under the weaker assumptions of Walley (1991) there is a partial weak order that is still transitive. Transitivity is also important in many applied situations. For example, clinical trials often employ the use of active-controls when investigating experimental treatments, as described by the International Conference on Harmonisation (2000). In an active-control trial, new experimental treatments are compared with those that have historically been shown to be efficacious over placebo and are currently in use. It is then common for regulatory agencies such as the US Food and Drug Administration to implicitly assume transitivity when assessing whether the experimental treatment should be approved for use: the trial results of the experimental treatment versus active-control are implicitly assumed to reflect the relationship between the efficacy of the experimental treatment versus placebo.

Given the theoretical and practical importance of transitivity, we seek to characterise the class of useful two-sample tests that maintain a proper ordering of distribution functions. The remainder of the manuscript is organised as follows: In Section 2 we define some useful terminology and the notation used throughout. Section 3 presents the main result of the paper, that under mild conditions a transitive test must be a test for a real-valued univariate parameter. Technically, our result is an extension of Debreu's theorems on preference relations and utilities (Debreu, 1954), but the application to statistical testing is new, and as far as we are aware the result is not a special case of any existing representation result. We conclude with a brief discussion of the theoretical and real-world importance of the scientific community's awareness of transitive inferential procedures.

2. Notation and definitions

We will say that a double one-sided test $U(\mathbb{F}, \mathbb{G})$ is *a test for* $T(\cdot)$ if for every level α it has two rejection regions: one satisfying $T(\mathbb{F}) < T(\mathbb{G})$ and the other satisfying $T(\mathbb{F}) > T(\mathbb{G})$. For example, the Welch two-sample t-test and Student's t-test are both tests for the mean and Student's t-test is a test for the mean that is asymptotically unbiased if the variances of F and G are equal. Similarly, Mood's median test is a test for the median (Brown and Mood, 1950), Bartlett's F test is a test for the variance (cf. Snedecor and Cochran, 1989), and Levene's test is for the mean absolute deviation from the mean (Levene, 1960). Clearly a test for a real-valued T must be transitive, and in the next section we will show that the reverse is almost always true, and is true for most useful tests. That is, under fairly weak conditions we will show that a transitive test must be a test for a real-valued statistic T.

Notation is simplified by assuming that the null value of the test $U(\mathbb{F}, \mathbb{F})$ is always zero by convention. Clearly this is no loss of generality. For example, the Mann–Whitney U-statistic on samples of size m and n has null value $mn/2$. However we can replace U by $U - mn/2$ as the test statistic. We further assume $U(\mathbb{F}, \mathbb{G}) = -U(\mathbb{G}, \mathbb{F})$. Thus the testing procedure reveals $\mathbb{F}$ is "better" than $\mathbb{G}$ precisely when it is reveals that $\mathbb{G}$ is "worse" than $\mathbb{F}$. Finally, by convention we write $\mathbb{F} \preceq \mathbb{G}$ when $U(\mathbb{F}, \mathbb{G}) \leq 0$, $\mathbb{F} \succeq \mathbb{G}$ when $U(\mathbb{F}, \mathbb{G}) \geq 0$, and $\mathbb{F} \approx \mathbb{G}$ when $U(\mathbb{F}, \mathbb{G}) = 0$.

Lastly, in our characterisation of tests it will be useful to consider the concept of a *total preorder*. Specifically, the relation $\preceq$ corresponding to test $U(\cdot, \cdot)$ is a total preorder if the following are satisfied: