Social Science Research 42 (2013) 482–495

Contents lists available at SciVerse ScienceDirect

Social Science Research

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

Permutation tests for goodness-of-fit testing of mathematical models to experimental data

M. Hamit Fişek *, Zeynep Barlas

Boğaziçi University, Istanbul, Turkey

ARTICLE INFO

Article history:
Received 23 February 2012
Revised 30 July 2012
Accepted 25 September 2012
Available online 4 October 2012

Keywords:
Social science standardized experimental settings
Goodness-of-fit tests
Resampling methods
The chi square test

ABSTRACT

This paper presents statistical procedures for improving the goodness-of-fit testing of theoretical models to data obtained from laboratory experiments. We use an experimental study in the expectation states research tradition which has been carried out in the "standardized experimental situation" associated with the program to illustrate the application of our procedures. We briefly review the expectation states research program and the fundamentals of resampling statistics as we develop our procedures in the resampling context. The first procedure we develop is a modification of the chi-square test which has been the primary statistical tool for assessing goodness of fit in the EST research program, but has problems associated with its use. We discuss these problems and suggest a procedure to overcome them. The second procedure we present, the "Average Absolute Deviation" test, is a new test and is proposed as an alternative to the chi square test, as being simpler and more informative. The third and fourth procedures are permutation versions of Jonckheere's test for ordered alternatives, and Kendall's tau_b, a rank order correlation coefficient. The fifth procedure is a new rank order goodness-of-fit test, which we call the "Deviation from Ideal Ranking" index, which we believe may be more useful than other rank order tests for assessing goodness-of-fit of models to experimental data. The application of these procedures to the sample data is illustrated in detail. We then present another laboratory study from an experimental paradigm different from the expectation states paradigm – the "network exchange" paradigm, and describe how our procedures may be applied to this data set.

1. Introduction

In this paper we present four permutation tests (Good, 2005) for assessing the goodness of fit of mathematical models to relevant experimental data. The development of these tests has been motivated by research in the expectation states tradition, and we shall choose our main illustrative examples from research in that tradition. However the applicability of these tests is certainly not limited to expectation states research: They can be applied in any research situation where an experiment with a number of conditions is carried out, and there is a mathematical model which predicts the values of a criterion variable for each condition depending on the manipulations involved in, and the subject population of, the particular condition. The basic question in such situations is whether the values of the criterion variable predicted by the model match the values observed in the experimental data. The tests we present are intended to answer this question. After presenting the tests and demonstrating their use on the expectation states data, we will describe how these tests can be applied to data from different standardized experimental situations.

* Corresponding author. Address: Dept. of Psychology, Boğaziçi University, Bebek, Istanbul 34342, Turkey. Fax: +90 212 287 24 72.
E-mail address: fisek@boun.edu.tr (M.H. Fişek).

0049-089X/$ - see front matter © 2012 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.ssresearch.2012.09.010

The first of these tests is an adaptation of the χ^2 test to the permutation testing framework. The χ^2 test is currently the most commonly used procedure in expectation states research, but it does involve problems which we will later discuss. We call the new procedure the "Monte Carlo Chi Square", and note that it is appropriate for experimental designs where the criterion variable is a "count" variable.

The second procedure which we call the "Average Absolute Deviation" (AAD)[1] test is more general than the χ^2 test in that its use is not limited to count data but can be used in any experimental situation if the criterion variable is metric.

The following three procedures are rank order procedures for use in situations where the model being tested does not provide metric predictions for the experimental conditions, but predicts the rank order of the conditions in terms of a statistic, most commonly the condition mean. The first two of these procedures are simply permutation versions of well-established tests, the first being Jonckheere's test for ordered alternatives (Jonckheere, 1954; Ager and Brent, 1978; Siegel and Castellan, 1988), and the second being Kendall's, tau_b (Gibbons, 1993; Higgins, 2004). The third is a new statistic which we call the "Deviation from Ideal Ranking" (DIR) statistic. As this is the first presentation of this statistic it is illustrated in some detail.

In this paper we present a brief introduction to resampling methods, followed by a brief introduction to expectation states research. The experimental study, "Behavior, Expectations, and Status" by Webster and Rashotte (2010) which provides the data for illustrating the use of our procedures is presented; followed by the presentation of the actual applications of the procedures to the sample data set.

Next we introduce two experimental studies in standardized experimental situations other than the expectation states standardized situation: One from the network exchange paradigm and one from the voting and agenda setting paradigm; we will explain how we selected these studies as we present them. What we do with them is to describe how our procedures can be applied to the data sets of these studies, and suggest what benefits will accrue from applying our procedures, compared to the standard procedures used. Unfortunately we cannot actually perform these analyses, because resampling statistics cannot be computed from summary measures as commonly reported in research articles, but only from data at the unit of analysis level.

We conclude with a few general remarks.

2. An overview of resampling methods

Resampling methods were introduced by a number of statisticians including Fisher (1935 as noted in Mielke and Berry, 2007) but were not widely adopted by practitioners. Interest in resampling methods has increased over time among statisticians (Good, 2006) and applications have also increased in the natural sciences but this tendency has not spread to sociology or the social sciences in general. Certainly there are no introductory statistics text books which take up resampling methods known to the current authors[2], even though their application requires lesser mathematical sophistication than the usual parametric methods.

The fundamental idea of resampling methods is to obtain the sampling distribution for a test statistic "empirically" rather than using theoretical distributions such as the normal, the *F*, the Chi Square, and the like, by taking samples from the actual research sample. "Bootstrap", "jackknife", and "permutation" are terms which name different methods of drawing samples from the original sample. The test statistic is computed for each of the "resamples" yielding the sampling distribution. The test statistic for the actual sample can then be located on the sampling distribution, and it can be determined whether it falls in a pre-established significance region.

The obvious and very important advantage of resampling statistics over parametric statistics is that they involve no assumptions (Good, 2005, 2006). Social science researchers commonly perform statistical tests without being sure that their situations meet the assumptions of the test being applied, or even knowing that the assumptions of the test are not met, but having no alternative test to apply. This advantage resampling methods offer is very important for the social science researcher.

The disadvantage of resampling methods is that they are very computation heavy, and historically this has been the reason for their being neglected (Good, 2006). However the current availability of computational power makes this disadvantage practically irrelevant and the advent of Monte Carlo procedures solves the problem by actually reducing the amount of computation necessary.

The permutation technique is the resampling type which is most suited to analyzing data from multi-condition experiments, and therefore the most appropriate for our purposes. Ideally the sampling distribution of the test statistic is obtained by computing the test statistic for each of all possible permutations of the subjects over the conditions of the experiment. A practical problem is that the number of permutations can be very large: an experiment with 5 conditions with 20 subjects in each condition has about 10^{66} permutations. Therefore Monte Carlo methods have to be used and a number of permutations have to be randomly selected. The number to be sampled is determined by the required precision of results. Our experimentation showed that 5000 replications gave results stable approximately to the third decimal place, and that is the number we have used in all analyses.

[1] We wanted to name this statistic, MAD for "mean absolute deviation", only to discover the acronym is already in use for "median absolute deviation" (Huber and Ronchetti, 2009).

[2] One exception to this statement is Fisher's Exact Test, which is a resampling procedure in that it depends on the exact enumeration of the sampling distribution but does not use Monte Carlo methods as it is restricted to very small samples. It is generally not recognized as a resampling procedure.