Available online at www.sciencedirect.com
ScienceDirect
Food Quality and Preference 18 (2007) 938–948
Food Quality and Preference
www.elsevier.com/locate/foodqual

Easy and powerful analysis of replicated paired preference tests using the χ^2 test ☆

Michael Meyners *

Givaudan, Huizerstraatweg 28, 1411 GP Naarden, The Netherlands

Received 16 June 2006; received in revised form 26 February 2007; accepted 11 March 2007
Available online 20 March 2007

Abstract

Assessor heterogeneity in replicated paired preference testing might mislead to the conclusion that there are no product differences at all, in particular if two equally sized consumer segments with opposed preferences occur. Different parametric approaches to deal with heterogeneity have been proposed, one of which is fitting a Beta-binomial model. Alternatively, we propose to use the ordinary χ^2-goodness-of-fit test to globally test for product differences. Examples are used to illustrate how the intermediate results of this test offer additional insight into the data and allow identifying possible consumer segments. Simulations show that the χ^2 test is more powerful than both the ordinary binomial and the Beta-binomial test, even if the data are truly Beta-binomially distributed. As the χ^2 approximation is liberal in many settings, we recommend using the Monte-Carlo simulation-based version of this test. We can easily perform this by using the open source software R.

Keywords: Replicated paired preference tests; Assessor heterogeneity; Consumer segmentation; χ^2-goodness-of-fit test

1. Introduction

Paired preference testing is frequently used to evaluate possible consumer preferences with respect to new or modified products. In its simplest form, consumers are asked to choose the preferred product out of two that are presented simultaneously. Usually, giving no preference is not allowed, such that the consumer have to choose one product at random if (they think) they do not have any preference. Pros and cons of this approach can be found in many textbooks (e.g., Lawless & Heymann, 1998) and will not be discussed here any further.

Consumers are questioned only once in most field applications. In order to obtain reliable information, there should be diversity in the sample of representative consumers, and this sample should be sufficiently large. An alternative approach might require consumers to evaluate the same two products repeatedly. This procedure could reduce the number of consumers involved, while difficulties might arise in ensuring that all participants re-do the test if the replicates cannot be performed immediately one after another. Although using a smaller number of different consumers might render the experiment less representative of the target population, individual preferences and assessor heterogeneity cannot be studied by means of unreplicated tests. Assume that one half of the target population has a strong preference for product A, while the other half has a strong preference for B. The expected test result of an unreplicated paired preference test in this case is exactly the same as if none of the consumers had any preference at all, or if consumers were not even able to tell the difference between the samples. Hence, the conclusion from such a study would be either that the products are almost identical, or that they are at least equally preferred. While the latter interpretation is not exactly wrong, it misses important information about the two consumer segments if the

☆ The work reported in this paper was done within Quest International Naarden.

* Tel.: +31 35 699 2349; fax: +31 35 699 5697.
E-mail addresses: michael.meyners@givaudan.com, m.meyners@web.de.

0950-3293/$ - see front matter
doi:10.1016/j.foodqual.2007.03.002

true preferences are indeed opposed to each other in two equally large parts of the population. Even if the population is unequally divided, knowledge about the segments can be very important. To illustrate this, assume that one ingredient in a product is to be replaced. It might occur that 70% of the population actually prefers the new product, but would still be pleased with the old one. The remaining 30% prefer the product as it is and strongly reject the new one, i.e. they would actually stop buying the product. It might then be better to leave the product unchanged, despite of the 70:30 preference for the new product. Even though replicated paired preference tests cannot reveal such detailed information, if analyzed appropriately they might give indications about possible consumer segments that need to be evaluated further.

Numerous approaches have been considered for the analysis of replicated paired preference tests. Many of them are suggested likewise for the analysis of replicated difference tests. These approaches include pre-testing for dependency of judgments (Smith, 1981), different generalized linear models (Hunter, Piggott, & Lee, 2000; see also Cochrane, Dubnicka, & Loughin, 2005, and the references therein), mixed models (Priso, Danzart, & Hossenlopp, 1994), Beta-binomial (BB) models (Ennis & Bi, 1998; Harries & Smith, 1982), and a binomial approach corrected for heterogeneity (Brockhoff & Schlich, 1998). We argue elsewhere (Meyners, 2007) that the analysis of two-sided preference and one-sided difference tests usually requires different approaches, as only chance-corrected models (cf. Brockhoff, 2003) are conceptually reasonable for the analysis of most one-sided difference tests. An exception is given by the ordinary binomial test, which is a valid test for either replicated preference or difference tests, given that the experiment is properly designed and performed (Kunert & Meyners, 1999). However, the power of this test decreases and the variability of the respective parameter estimates increases if evaluations of independent assessors are replaced by replications without appropriately increasing the total number of experiments (Meyners & Brockhoff, 2003). As shown in the previous paragraph, systematic but opposing preferences for the products will cancel each other more or less completely in certain situations. The power of the binomial test will be particularly poor then.

According to the authors cited above and the examples given in their respective publications, all suggested methods have in common that they are only used to correct for heterogeneity. Even though theoretically possible, explicit tests and estimates of the parameters representing heterogeneity are rarely performed, not to speak of any interpretation thereof. This becomes most evident in the comparison of analyzing methods presented by Cochrane, Dubnicka, and Loughin (2005). For replicated paired preference tests, the authors base their comparison of several approaches solely on the respective rejection rates of the hypothesis that the mean preference rate is the same for both products under consideration. With respect to power, the differences between the methods were minor, and the normal approximation of the ordinary binomial test performed quite well. This latter approach proved to be liberal if both products were indeed equally preferred (with different levels of heterogeneity), but this problem could be easily fixed by using the exact binomial test instead of the normal approximation. Hence, if only the average preference rates of the products were of interest, the binomial test would be the method of choice, even more as it does not rely on any distributional assumptions that cannot be warranted by means of the experimental design. However, if heterogeneity is indeed of concern, it should not only be modeled, but also be tested and interpreted. Apparent differences between assessors necessarily imply that the products are indeed different from each other and that not all assessors equally prefer any product. This holds even if the mean preference rates are more or less identical, as in the example given earlier in this section. This is shown in more detail for replicated difference tests by Meyners (2007), but it holds all the same for replicated paired preference tests.

Taking these considerations into account, it seems reasonable to analyze the results of a replicated paired preference test by means of either an omnibus test for deviations from the null distribution, or by means of a multi-parameter approach that includes both the mean preference rate and the assessor heterogeneity, and which gives interpretable results for these parameters. In the following, we will show that the well-known χ^2-goodness-of-fit test is a good alternative to other methods, while it does not depend on any parametric model assumptions. We describe how this test should be used and how the intermediate results can be interpreted in order to derive additional information for further evaluation of the products. The code for an easy implementation in the open source software R is provided. Examples from the literature are re-analyzed and the results are compared to those of the respective original analysis. We show that possibly important information was neglected in these examples by focusing on the mean preference rates only. Finally, a simulation study compares this method with the ordinary binomial test as well as with the simultaneous test of both parameters from the BB model. Even though the χ^2 test is widely used in other applications, to our knowledge its application in this way to replicated preference or difference tests has not yet been described.

2. Application of the χ^2-goodness-of-fit test

We will describe the χ^2 test itself only briefly as it is a well-known test discussed in many textbooks on statistics (e.g. Zar, 1996). Instead, we will pay more attention to the rationale in this particular application and to possible interpretations of the intermediate results.

Assume that n assessors consider the same pair of products k-fold. The null hypothesis of a paired preference test states that there are no differences in preference between the products A and B under consideration. As the assessors