Available online at www.sciencedirect.com
ScienceDirect
Food Quality and Preference 19 (2008) 292–305
Food Quality and Preference
www.elsevier.com/locate/foodqual

Hierarchical Bayesian analysis of true discrimination rates in replicated triangle tests

Kees Duineveld [a,1], Michael Meyners [b,*,2]

[a] Givaudan, Huizerstraatweg 28, 1411 GP Naarden, The Netherlands
[b] Holzbaumweg 34, 49740 Haselünne, Germany

Received 11 May 2007; received in revised form 5 October 2007; accepted 21 October 2007
Available online 30 October 2007

Abstract

Several methods for testing the null hypothesis in replicated triangle tests are known and frequently used. Deriving further reliable information about the distribution of discrimination rates can be more difficult due to possible assessor heterogeneity, though. In contrast to unreplicated tests, it cannot be assumed that an assessor has truly discriminated either never or always. To overcome this problem, we use a Bayesian hierarchical model to estimate the distribution of the true discrimination rates. For the calculations, we apply a Monte Carlo Markov Chain (MCMC) sampler. An example employing 30 assessors with 3–10 replications each indicates that roughly about 1/3 of the panelists is not able to differentiate between the products, while another third has most likely a noteworthy discrimination probability. The last third is somewhere in between, with some chance that they can hardly discriminate, but also some chance that they indeed have a positive discrimination rate between 0.1 and 0.5. The distribution of these probabilities as well as of the mean discrimination rate can be estimated from the model. In our example, the average true discrimination rate is estimated as 0.32 with a corresponding 95% confidence interval of [0.23; 0.41]. This suggests that the average probability for the difference between the products being truly perceived lies between 23% and 41%. General results indicate that at least 5 or 6 replications are needed to reasonably approximate individual panelists' behaviour in triangle tests.

Keywords: Bayesian analysis; Hierarchical models; Replicated triangle test; Heterogeneity; Beta-binomial model; True discrimination rate

1. Introduction

In one-sided difference testing, replications of the test by the same assessors are sometimes considered to extend the data, for example in order to increase the power without increasing the number of assessors (see, e.g., Brockhoff, 2003; Meyners & Brockhoff, 2003). If no product differences are present, this is no problem at all, while it complicates the analysis if the products truly differ. In the latter case, the analyst usually wants to investigate the overall or individual discrimination rates, while such investigations need to take possible heterogeneity into account. Technically speaking, heterogeneity implies different levels of dependency of trials within and between assessors, respectively. In what follows, we confine ourselves to the triangle test as the probably most frequently used one-sided difference test. The transfer to other tests is straightforward.

For an unreplicated triangle test, the binomial test can be used to test whether the proportion correct is 1/3. In addition to the proportion correct, measures such as d' (distance between products) and proportion true discriminators (proportion of correct answers exceeding the expected proportion of 1/3) can be determined. The details are well-known and can be found in many textbooks (e.g. Lawless & Heymann, 1998). First approaches to show product equality up to a certain margin use the so-called

* Corresponding author. Tel.: +31 36 529 81 50.
E-mail addresses: kees.duineveld@givaudan.com (K. Duineveld), m.meyners@web.de (M. Meyners).
[1] Tel.: +31 35 699 2460; fax: +31 35 699 5697.
[2] This author undertook parts of the work while working at Givaudan.

0950-3293/$ - see front matter
doi:10.1016/j.foodqual.2007.10.005

power approach. The power of the triangle test is examined by Ennis (1993) for given values of d' and by Schlich (1993) for given proportions of true discriminators. In addition, MacRae (1995) derives confidence levels for the proportion of true discriminators. Bi (2005) reverses the test problem, proposing an equivalence test to prove that the proportion of discriminators does not exceed a certain value. This approach has been extended by Meyners (2007a), introducing the least equivalent allowable difference (LEAD). Products will be claimed to be equal in case the difference does not exceed a certain margin, and the LEAD gives the minimal margin such that the equivalence test still yields statistical significance.

At the same time, Bayesian methods get increasingly important. The basic theorem is known as Bayes' formula, and it essentially states that the conditional probability of A given B, $P(A|B)$, is proportional to the product $P(B|A)^*P(A)$ for arbitrary events A and B. This formula applies accordingly to probability densities, and it is used to combine the information from an observed data set with some prior information $P(A)$. Note that the prior can (and frequently will) be more or less non-informative.

Until the last decade of the 20th century, it was hardly possible to use Bayes' formula except for some special cases where the corresponding equations can be solved analytically. Using, for example, a Beta distribution as a prior for a binomial model results in an analytically accessible model; this can be considered as a main feature supporting the popularity of the Beta-binomial (BB) model (Skellam, 1948) in many areas. The use of alternative priors was very difficult or even impossible to deal with. However, the advance of computing power and new developments resulted in methods such as Monte Carlo Markov Chain (MCMC), giving a numerical approximation of the posterior distribution (see, e.g., Casella & George, 1992; Metropolis & Ulam, 1950). Development of these methods is still ongoing with high speed, and many models are considered which would be difficult to implement in classical, frequentistic statistics. Elaborate introductions into Bayesian methods and examples from different branches of science can be found in many papers (e.g. Casella & George, 1992; Smith & Roberts, 1993) and textbooks (e.g. Gelman, Carlin, Stern, & Rubin, 2004).

Numerous methods have been proposed for the analysis of replicated triangle tests. A BB model (Skellam, 1948) is proposed by Harries and Smith (1982), to which Ennis and Bi (1998) add power calculations. Priso, Danzart, and Hossenlopp (1994) propose a combined hypothesis test for the proportion correct and the dispersion, while Brockhoff and Schlich (1998) use an overdispersion parameter to correct the number of observations for heterogeneity, followed by a binomial test. Kunert and Meyners (1999) show that the naïve binomial test can be used to test the null hypothesis of product equality without violating the given level. However, this approach does not support additional interpretation beyond the global test; heterogeneity, which might occur in case the products are truly different, is not taken into account. Kunert (2001) extends the approach by allowing individual discrimination rates. He derives a confidence interval for the overall discrimination probability using the central limit theorem. Hunter, Piggott, and Lee (2000) consider a generalized linear model, in which a large residual deviance of the model indicates assessor heterogeneity. They also correct the confidence intervals for the proportion of correct answers by means of the mean deviance in case of apparent heterogeneity. Brockhoff (2003) argues that both the BB model and the generalized linear model suffer a fundamental flaw, namely that fitted values with a proportion correct under 1/3 are possible (see also Kunert & Meyners, 1999; Meyners, 2007b). He therefore derives chance-corrected versions of both models, forcing the estimated proportion of correct answers to lie between 1/3 and 1. Power calculations are given for these models as well. With respect to the reliability of the estimates and the power of the test, the impact of the number of replications compared to the total number of experiments in replicated tests is extensively discussed by Meyners and Brockhoff (2003). Meyners (2007b) shows that assessor heterogeneity on its own implies product differences and compares the standard BB approach with the chance-corrected one by means of simulations. For replicated two-sided tests, Meyners (2007c) shows that the ordinary χ^2-goodness-of-fit test is a powerful alternative to other models in order to detect and interpret product differences.

Bi (2003) was probably the first to use a Bayesian framework for the analysis of replicated triangle tests, deriving posterior confidence intervals on the proportion correct and using so-called Bayes factors to decide on alternative hypotheses. More recently, the same author proposed a Bayesian approach for unreplicated tests (Bi, 2007), using the idea of the chance-corrected BB model as given by Brockhoff (2003). Carbonell, Carbonell, and Izquierdo (2007) used Bayes' rule to obtain a distribution for the proportion of true discriminators in unreplicated tests. This approach was generalised to obtain a distribution of true discrimination rates in replicated tests by Meyners and Duineveld (2008). The present paper can be seen as a further generalisation of this approach.

Apart from statistical significance testing and the interpretation of d', most methods proposed yet do not allow assessing the distribution of discrimination probabilities or individual discrimination rates. To extend the knowledge, we use a Bayesian hierarchical model which is related to the chance-corrected BB model of Brockhoff (2003). However, there are a number of differences. In our model, the variability of individual true discrimination rates is taken into account. Also, panelists with less than 1/3 correct answers are handled differently, as it is perfectly reasonably that such a panelist truly discriminated once or twice, but was unlucky with the remaining tests. The respective probabilities are derived using the approach of Carbonell et al. (2007), applying their approach to each assessor individually. We explicitly model the panel behav-