Deriving a lower bound for the proportion of perceivers in replicated difference tests by means of multiple test theory

Michael Meyners
Fachbereich Statistik, Universität Dortmund, D-44221 Dortmund, Germany
Tel.: +49 231 755 3181 Fax: +49 231 755 3454
E-Mail: michael.meyners@udo.edu

Abstract

Analyzing repeated difference tests aims in significance testing for differences as well as in estimating the mean discrimination ability of the consumers. In addition to the average success probability, the proportion of consumers that may detect the difference between two products and therefore account for any increase of this probability is of interest. While some authors address the first two goals, for the latter one only an estimator directly linked to the average probability seems to be used. However, this may lead to unreasonable results. Therefore we propose a new approach basing on multiple test theory. We define a suitable set of hypotheses that is closed unter intersection. From this, we derive a series of hypotheses that may be subsequently tested while the overall significance level will not be violated. By means of this procedure we may determine a minimal number of assessors that must have perceived the difference between the products at least once in a while. From this, we can find a conservative lower bound for the proportion of perceivers within the consumers. In several examples, we give some insight into the properties of this new method and show that the knowledge about this lower bound might indeed be valuable for the investigator. Finally, an adaption of this approach for similarity tests will be proposed.

Key words: Replicated difference tests, perceivers, estimation, multiple tests

Introduction

We consider repeated difference tests which are intended, e. g., to determine whether or not differences between two products of the same kind exist. With it, we are usually interested in differences with respect to taste, smell or appearance. Mostly non-replicated tests are investigated in the area of sensory analysis, where usually consumers are considered. A well-known difference test is the so-called triangle test. For this, three samples of the products under consideration are used: two of product A, say, and one of another product B, say. These samples are arranged in triangular form and presented to the assessors. Those are asked to assess the samples and identify the odd sample, i. e. the one that differs from the other ones. Details concerning the design of the experiment are beyond the scope of this paper, but it is easy to see that the probability of just guessing right by chance is equal to $\frac{1}{3}$. This means that whenever the judges cannot find the odd sample due to product differences or an inappropriate design and whenever the order of presentation is chosen at random from one of the six possibilities AAB, ABA, BAA, ABB, BAB and BBA, the success probability is $\frac{1}{3}$. For details we refer to the paper by Kunert and Meyners (1999).

To analyze the outcomes of such an experiment, a simple binomial test ist used. In a triangle test, we are concerned with the validity of the null hypothesis H_0: "There are no differences between the products". This, of course, is equivalent to testing H_0: "$\pi = \frac{1}{3}$" versus H_1: "$\pi > \frac{1}{3}$", while π is the mean success probability within the set of assessors. We may confine ourselves to this one sided test since the success probability may not fall below $\frac{1}{3}$ for any assessor in case the experiment has been properly designed (Kunert and Meyners, 1999).

If the product differences are large, a small number of assessors may be sufficient to prove that there are indeed differences. However, in practical applications this kind of experiment is rather used when the expected differences are quite small. Therefore, due to intra product variations caused by, e. g., different stocks, the assessors may fail to identify the odd sample once in a while, even if there are differences present. Therefore, a large number of experiments has to be conducted to find the differences between the products in case there are any. With it, a large number of consumers has to be tested. This, of course, is very time consuming as well as expensive. Hence considerations are given to testing the same consumers repeatedly. Under proper randomization it can be shown that this does not influence the analysis of the data with respect to the significance test (Kunert and Meyners, 1999). Hence we might still use the binomial test with $\sum_{i=1}^{n} k_i$ observations, while k_i is the number of replications of assessor i.

In general we are interested not only in knowing whether or not a difference between the products exists, but also rather in estimating the mean success probability above chance, for which an easily determined estimator is common practice. Of additional interest for the investigator is to judge upon the number of perceivers within the consumers, i. e. those consumers that might detect the difference at least once in a while. It may matter whether an overall 50% above chance success probability is derived from all consumers detecting the difference in every second trial, or from one half of the consumers that will almost surely detect the difference while the other half never does. Assuming the latter case, an estimator of the number of perceivers can be easily derived. However, this might lead to unreasonable results as will be shown in what follows. The aim of this paper is to develop a method in order to derive a lower bound for the proportion of perceivers by means of multiple test theory.

Model

For each assessor $i \in \{1, ..., n\}$, her/his success probability is π_0, say, in case he does not taste the difference in a particular experiment and therefore she/he has to guess which sample may be the right one. Contrariwise, the respective probability is 1 in case the assessor perceives the difference in this experiment. For the triangle test, we find $\pi_0 = \frac{1}{3}$, while we may have different values of π_0 for other difference tests like, e. g., the duo-trio test. Hence random variable representing the overall success probability Π_i, say, of assessor i is given by

$$\Pi_i = \pi_0(1 - \Theta_i) + \Theta_i = \pi_0 + (1 - \pi_0)\Theta_i \ , \tag{1}$$

where Θ_i is the random variable that represents the individual perceiving probability of assessor i with values in $[0, 1]$. For a non-perceiver, Θ_i is almost sure equal to zero and hence the expectation $\theta_i := \mathbf{E}(\Theta_i)$ is equal to zero. On the other hand, if $\mathbf{E}(\Theta_i) > 0$ and hence $\pi_i := \mathbf{E}(\Pi_i) > \pi_0$, we call assessor i a perceiver. Note that this is the most general model for difference tests, since all restrictions are derived from properties that can be assured by means of the design of the experiment (Kunert and Meyners 1999). From this formula, we will derive a lower bound for the proportion of perceivers by means of multiple test theory.

Different models have been proposed for the individual success probabilities. Besides the one given here, the Beta-binomial model has been frequently addressed (Ennis and Bi 1998, Bi and Ennis 1999a, b). This model has a main drawback since it allows the success probability to fall below π_0, which may only occur with an inappropriate design of the experiment (Kunert and Meyners 1999). A corrected version of the Beta-binomial model has been proposed by Brockhoff (2002) and examined by Meyners and Brockhoff (2002) as well. Assuming this model, the goal would rather be to estimate the parameters of the distribution, which can be done by means of, e. g., Maximum-Likelihood estimators. However, we do not aim at this for the following reasons. First of all, no textual justification of the individual success rates being (corrected) Beta-distributed can be seen. Instead, the main rational behind this choice seem to be the easily determined statistical properties. Secondly, our approach does not rely at all on the distribution assumption in model (1) for Θ_i respectively Π_i, but only on the definition of a perceiver by means of $\theta_i > 0$ respectively $\pi_i > \pi_0$. Therefore our approach gives a non-parametric lower bound for the proportion of perceivers. As it has been mentioned above, the model assumptions are completely general such that non-robustness of the approach cannot pose any problems.

Closed set of hypotheses under intersection

With respect to model (1), we may re-write the test problem in a more statistical notation as

$$H_0 : \pi_i = \pi_0 \quad \forall \quad i \in \{1, ..., n\}$$

versus

$$H_1 : \exists \quad i \in \{1, ..., n\} : \pi_i > \pi_0$$

respectively

$$H_0 : \theta_i = 0 \quad \forall \quad i \in \{1, ..., n\}$$

versus

$$H_1 : \exists \quad i \in \{1, ..., n\} : \theta_i > 0.$$

For simplicity of notation, in what follows we confine ourselves to the former representation of the problem. Let the random variable X_i denote the number of correct answers of assessor i. Assuming a proper design of the experiment, we may assume the outcomes of each test to be independent, and therefore the distribution of X_i is binomial with parameter π_i and k_i observations, shortly $L(X_i) = \mathbf{B}(k_i, \pi_i)$.

Let $H = \{H_0^P : P \subset \{1, ..., n\}\}$ be the set of the following subsets of null hypotheses:

$$H_0^P : \pi_i = \pi_0 \quad \forall \quad i \in P.$$

Since $P_1 \cap P_2 \subset \{1, ..., n\}$ for any subsets P_1 and P_2 of $\{1, ..., n\}$, we find that $H_0^{P_1} \cap H_0^{P_2} \in H$ for any subsets of hypotheses $H_0^{P_1}$ and $H_0^{P_2} \in H$. Therefore this set of subsets of hypotheses is obviously closed under intersection.

We will now use this property of the set of hypotheses to address our purpose, i. e. to estimate the minimal number of perceivers.

Multiple test procedure

Our main concern is to test the following sets of hypotheses:

$$H_0^l : \pi_i = \pi_0 \text{ for at least } n - l + 1 \text{ assessors}$$

versus

$$H_1^l : \pi_i > \pi_0 \text{ for at least } l \text{ assessors,}$$

where $l \in \{1, ..., n\}$. However, these hypotheses cannot be tested directly without violation of the significance level α, say. Therefore we re-write and enlarge the set of subsets of hypotheses to a set that is closed unter intersection, namely the one presented in the section before. We may now equivalently write H_0^l as

$$H_0^l : \quad \exists \quad P_l \subset \{1, ..., n\} \quad \text{with} \quad |P_l| = n - l + 1 : \pi_i = \pi_0 \quad \forall \quad i \in P_l.$$

Marcus, Peritz and Gabriel (1976) propose a procedure that allows the performance of a series of subsequent level-α tests for these hypotheses while the overall probability of making a type I error will not exceed α. For this series, it only has to be guaranteed that all tested hypotheses are from the set given above – which are actually all hypotheses of interest within this context – and that we test a subset of hypotheses $H_0^{P_b}$ only if all subsets of hypotheses $H_0^{P_a} \in H$, that are included in $H_0^{P_b}$, have been tested and rejected before. Here, these subsets of hypotheses are defined by $P_b \subset P_a$, since, e. g., the null hypothesis that "$\pi_i = \pi_0$ for all $i \in \{1, ..., n\}$" is obviously a particular case of the subset "$\pi_i = \pi_0$ for all $i \in \{1, ..., n-1\}$". Hence, following Marcus *et al.* (1976), we may test "$\pi_i = \pi_0$ for all $i \in P_b$" versus "$\pi_i > \pi_0$ for at least one $i \in P_b$" at level α whenever we could reject all subsets of hypotheses "$\pi_i = \pi_0$ for all $i \in P_a$". In these cases we accepted "$\pi_i > \pi_0$ for at least one $i \in P_a$" for all supersets $P_a \supset P_b$.

Applying this theory to the difference tests under consideration, from the technical restrictions given above we find that the following subsequent testing procedure will restrain the type I error rate at level α:

step 1: Test the set of null hypotheses "P_n: $\pi_i = \pi_0 \;\; \forall \;\; i \in \{1, ..., n\}$" versus "$\exists \;\; i \in \{1, ..., n\} : \pi_i > \pi_0$" at level α.

step 2: If the set of null hypotheses P_n could be rejected, test now all subsets of null hypotheses "$\pi_i = \pi_0 \;\; \forall \;\; i \in P_{n-1}$" versus "$\exists \;\; i \in P_{n-1} : \pi_i > \pi_0$" at level α, while P_{n-1} is an arbitrary subset of $\{1, ..., n\}$ of length $n-1$.

step 3: If there exist subsets P_{n-2} of $\{1, ..., n\}$ of length $n-2$ such that all subsets of null hypotheses P_{n-1} with $P_{n-1} \supset P_{n-2}$ have been rejected in the former step, test these subsets P_{n-2} at level α.

...

step 4: If there exist subsets P_{n-l+1} of $\{1, ..., n\}$ of length $n-l+1$ such that all subsets of null hypotheses P_{n-l+2} with $P_{n-l+2} \supset P_{n-l+1}$ have been rejected in the former step, test these subsets P_{n-l+1} at level α.

...

In each step l, in case there is no set of subsets of hypotheses P_{n-l+1} such that the respective condition is fulfilled, the procedure has to be stopped. Hence, whenever not all supersets of hypotheses including P_{n-l+1} could be rejected, no additional test should be performed anymore.

Still, for our purpose this procedure might be simplified in most cases. We now confine ourselves to the case where $k_i = k$, i. e. all assessors perform the difference test equally often, which usually holds whenever replicated difference tests are considered. Assuming a proper design of the experiment, the results of different assessors can be assumed to be independent. For each subset $P \subset \{1, ..., n\}$ we know that $L(\sum_{i\in P} X_i) = \mathbf{B}(\sum_{i\in P} k_i, \pi_0)$ whenever the corresponding null hypotheses hold, i. e. whenever $\pi_i = \pi_0$ for all $i \in P$. Note that this is the case in which no detectable product differences are given for the assessors within this experiment.

Hence the appropriate test in each step is the simple binomial test with parameter π_0 and $|P|k$ observations, while $|P|$ denotes the length of P. In step 1 we hence use nk observations. In case we can reject the set of null hypotheses, we accept that there is at least one assessor i_1, say, for whom $\pi_{i_1} > \pi_0$. We conclude that there is at least one perceiver out of n consumers within this panel. (Note that we are not interested in judging which assessors have perceived the difference, but only in the proportion of perceivers in all. The other problem would require a large number of replications k_i for each assessor and is usually not of great interest in practice.) Hence we know that we might go to step 2: for all subsets of length $n-1$ of assessors we test whether there is still at least one within each subset who might perceive the difference. Of course, knowing that in most of these subsets assessor i_1 will be included, it is not of much use testing all these hypotheses once again.

Remembering the test problem under consideration, in this second step we only want to know whether there is at least one perceiver in *each* subset of length $n-1$. In this step, we use the binomial distribution with $(n-1)k$ observations. As it is well known, we will reject the null hypothesis and assume the alternative whenever the observed total number of correct answers is too large. Thus we may confine ourselves to the case in which this number is the smallest within all cases. Obviously, this is the case in which the assessor with the most correct answers is removed from the complete set. In case there are several assessors with this number of successes, we have different possibilities. Fortunately this has no influence on the outcomes, i. e. it makes no difference which one will be removed. Let i_1 denote the assessor that is removed now.

For the next step, note that in case we reject the null hypothesis without the assessor mentioned before, we also would have rejected this hypothesis leaving out any other assessor. In all those cases, the observed test statistic would have been larger (or at least not smaller) while the critical value remains the same. Hence, we implicitly have tested all other hypotheses. This allows us to proceed with the next step according to Marcus *et al.* (1976) and to the procedure described above, namely to test the hypotheses with any subset of length $n-2$ of the assessors. Again, we may confine ourselves to the subset in which one of the assessors with the largest number of successes is removed and get to the next step, using subsets of length $n-3$.

Here, we have to pay attention to the inclusion condition of Marcus *et al.* (1976). From all subsets of length $n-2$, we explicitly considered only those determined from the subset of length $n-1$ chosen in step 2. The corresponding hypotheses, of course, could all be neglected, if the one considered in step 2 could be so, leaving out the most successful assessor. On the other hand, consider the subsets of length $n-2$ that are determined from any other subset of length $n-1$, leaving out any other assessor but i_1. The corresponding hypotheses imply that $\pi_{i_1} = \pi_0$, while this is not included in the respective hypothesis for the subset of length $n-2$ under consideration. Hence, these hypotheses are not included in the one considered here. Therefore it is not of any interest whether the hypotheses corresponding to the subsets of length $n-2$ that are *not* determined from the subset of length $n-1$ from step 2 could be rejected or not.

For the further steps, the following consideration is of additional importance: Consider that not i_1 has been removed in the first step but i_2, say. Then, in the second step we might remove i_1 which hence results in a subset of importance for the multiple testing procedure, but which has not been explicitly tested within this procedure. However, this subset is obviously identical with the one derived by removing i_1 first and then i_2, i. e. this subset has indeed been tested, even though derived in another way this time.

To summarize, without loss of generality, we assume the assessors to be numbered such that the one with the most successes is assessor n and the one with the least successes is assessor 1. For assessors with identical numbers, an arbitrary order might be chosen. If now $k_i = k$ for all i, we may use the following procedure:

step 1: Test the set of null hypotheses "P_n: $\pi_i = \pi_0 \ \ \forall \ \ i \in \{1, ..., n\}$" versus "$\exists \ \ i \in \{1, ..., n\} : \pi_i > \pi_0$" at level α.

step 2: If the set of null hypotheses in step 1 could be rejected, remove assessor n (i. e. the one with most successes) from the data set and test the set of null hypotheses "P_{n-1}: $\pi_i = \pi_0 \ \ \forall \ \ i \in \{1, ..., n-1\}$" versus "$\exists \ \ i \in$

$\{1, ..., n-1\} : \pi_i > \pi_0$" at level α.

step 3: If the set of null hypotheses in step 2 could be rejected, remove assessor $n-1$ (i. e. the one with the second most successes) from the data set and test the set of null hypotheses "P_{n-2}: $\pi_i = \pi_0 \quad \forall \quad i \in \{1, ..., n-2\}$" versus "$\exists \quad i \in \{1, ..., n-2\} : \pi_i > \pi_0$" at level α.

...

step l: If the set of null hypotheses in step $l-1$ could be rejected, remove assessor $n-l+2$ (i. e. the one with the $(l-2)$-most successes) from the data set and test the set of null hypotheses "P_{n-l+1}: $\pi_i = \pi_0 \quad \forall \quad i \in \{1, ..., n-l+1\}$" versus "$\exists \quad i \in \{1, ..., n-l+1\} : \pi_i > \pi_0$" at level α.

...

In each step l, the appropriate test is the binomial test using the binomial distribution with parameter π_0 and $(n-l+1)k$ observations. The procedure stops as soon as a set of null hypotheses cannot be rejected anymore. With it, we have a test procedure for the initial problem from the beginning of this chapter which restrains the type I error at level α.

To be precise, this procedure does not only hold the level α, but is conservative! Assume that $k = 3$ in a triangle test, i. e. $\pi_0 = \frac{1}{3}$. Then the probability that a single assessor succeeds thrice by pure guessing is given by $\frac{1}{27}$ and therefore smaller than, e. g., 5%. We would not expect a single assessor to give this results by chance only. On the other hand, if we have 30 assessors who are pure guessers, we would nevertheless expect one of those to give three correct answers. On the other hand, if a perceiver has an overall success probability of $\frac{1}{2}$, say, he might succeed once or twice only in three replications as well. Hence, if there are a lot of non-perceivers within the panel and only few perceivers with a moderate success probability, we will probably remove some of the non-perceivers from the data set instead of the perceivers, which is due to their larger success rate by chance. On the other hand, this cannot be helped, since we cannot judge by any means which particular assessors perceived a difference, therefore the application of this conservative procedure is essential to restrain the type I error rate at the given level α.

Finally, we may now go back to the initial test problem, namely to test whether or not there are at least l assessors out of n that perceive a difference at least once in a while. For these assessors, $\pi_i > \pi_0$ holds. Let the procedure stop in step l, i. e. the corresponding hypothesis cannot be rejected anymore. Thus, from the construction of the procedure described above, we have shown that there are at least $l-1$ perceivers at a significance level of α. Hence, a (conservative) lower bound for the proportion of perceivers within the panel is given by $\frac{l-1}{n}$.

Relation to the conventional approach

The most frequently used estimate for the average success probability above chance within the consumers is given by $\frac{\hat{\pi}-\pi_0}{1-\pi_0}$ whenever this value is positive. Here, $\hat{\pi} = \frac{x}{nk}$ is the ratio between the total number of successes x and the total number of assessments nk. If we are interested in the proportion of perceivers within the consumers δ, say, the same estimator is mainly used, i. e. the estimator

$$\hat{\delta} = \frac{\hat{\pi} - \pi_0}{1 - \pi_0}.$$

This only seems to be justified in case of an additional assumption: If we can assume that the success probability in each trial is either 1 or π_0 and remains the same for each assessor, this approach is reasonable and justified. For the interpretation, this means that whenever an assessor identifies the odd sample at least once, she/he will always identify it. However, this does not seem reasonable as far as the products do not differ too much from each other such that each assessor always correctly identifies the odd sample. Furthermore, the estimate and the assumption might contradict each other. For a triangle test, Meyners (2002) considers the case of $n = 20$ and $k = 3$ in which all assessors give 2 correct answers each. Hence the conventional estimate for the overall success probability above chance is given by $\frac{1}{2}$, which could be a reasonable value. Contrariwise, using this one for the proportion of perceivers as well given the conditions mentioned above, this does not make any sense: From the assumption there cannot be any perceiver at all – none of the assessors identified the odd sample in each trial! On the other hand, such outcomes would give strong hints that a large number of the assessors under consideration indeed perceived the difference between the products once in a while. The results are far beyond chance in case of product equality.

From the procedure proposed here, we would estimate a lower bound for the number of perceivers to be given by $\frac{18}{20} = 0.9$, such that we would conclude that most consumers will indeed detect the difference between the varieties of the product under consideration. With it, we would claim that they are perceivers. Noting that this is a lower bound, obviously the estimate given before does not seem realistic at all. Hence, an additional information is available from this procedure. In this extreme example, we would get a much better impression of the data from the lower bound than from the conventional estimate only.

Note that the procedure presented here guarantees that we will estimate the number of perceivers by a value not smaller than $\frac{1}{n}$ if and only if significant differences between the products by means of the binomial test with nk observations can be proven (cf. Kunert and Meyners, 1999). This fact represents the well known duality between statistical significance testing and confidence intervals – we hence may consider our value as a lower bound of a confidence interval for the proportion of perceivers. Using this method might only result in a lower bound, while the upper bound would be given by 1. Considering our procedure the other way round, we might also end up with an upper bound. However, for this we would have to consider the inverse test problems, i. e. for any $l \in \{1, ..., n\}$

$$H_0^l : \pi_i > \pi_0 \text{ for at least } n - l + 1 \text{ assessors}$$

versus

$$H_1^l : \pi_i = \pi_0 \text{ for at least } l \text{ assessors.}$$

The values of π_i cannot fall below π_0 whenever the experiment has been properly designed. Hence the test problem is equivalent to re-writing the alternative hypotheses as

$$H_1^l : \pi_i \leq \pi_0 \text{ for at least } l \text{ assessors,}$$

for which the appropriate statistical test is well known. In this case, we derive the number l the other way round, i. e. we remove those assessors that give the smallest number of successes. Using the same arguments as before, this will result in a lower bound for the number of non-perceivers respectively an upper bound for the number of perceivers. Adding an arbitrary value ρ to π_0, we may even confine ourselves on perceivers with a minimum success probability above chance, which is probably of more interest within applications. If we are interested in a 10% success probability above chance, we would therefore consider $\rho = (1 - \pi_0) * 0.1$ and consider the test problems

$$H_0^l : \pi_i > \pi_0 + \rho \text{ for at least } n - l + 1 \text{ assessors}$$

versus

$$H_1^l : \pi_i \leq \pi_0 + \rho \text{ for at least } l \text{ assessors}$$

for $l \in \{1, ..., n\}$. The non-perceivers then include those assessors with a slightly increased success probability up to $\pi_0 + \rho$ as well. If we only consider the upper bound, a lower one for the corresponding confidence interval would be zero. Deriving an upper bound may be of particular interest whenever we are interested in proving similarity in case no differences occurred. In this context it is common practice to consider differences only if they are larger than a pre-defined minimal effect size.

Note that we have to decide in advance about the confidence interval of interest. Considering an identical value of ρ to determine both the lower and the upper bound would result in a two-sided interval. Even though it is clear from the construction that at least either the lower bound is 0 or the upper bound is 1, we are not allowed to make the decision about which one to consider according on the outcomes. This would result in a two-sided confidence interval at level 2α! If we are interested in a two-sided one at level α, we have to perform the multiple test procedures at level $\frac{\alpha}{2}$.

Examples

In what follows, we give some more examples to illustrate the properties of the lower bound for the proportion δ under consideration. All these examples refer to the triangle test and some of them have been discussed in more detail by Meyners (2002). Here, we aim to give an impression of the properties of the newly proposed method with respect to some data given in the literature. For this issue, we generally assume a 5% significance level.

First of all, we consider three data sets that have been presented by Hunter, Piggott and Lee (2000). In each test, the number of replications was k=12. In the first trial, the number n of assessors was 30, whereas it was 24 in the second and 23 in the last set. The numbers of assessors giving x right answers, $x \in \{0, 1, ..., 12\}$, are given in table 1.

x	0	1	2	3	4	5	6	7	8	9	10	11	12	total
exp. 1	0	0	1	2	3	7	8	6	2	1	0	0	0	170
exp. 2	1	0	1	5	5	3	3	3	1	2	0	0	0	117
exp. 3	0	0	2	1	1	4	3	6	3	1	0	1	1	147

Table 1: Number of assessors with x right answers and total number of successes for the three experiments reported by Hunter *et. al.* (2000).

For the first data set, the average success probability above chance would be estimated by $\frac{3}{2}\left(\frac{170}{360}-\frac{1}{3}\right)\approx 0.21$ by means of the conventional approach. At a 5%-level we identify 13 perceivers, leading to $\hat{\delta}=\frac{13}{30}\approx 0.43$. Hence we conclude that at least 2 out of 5 consumers figure out the difference once in a while. Maybe we would not worry about a fifth of the consumers figuring out the difference always, but we might worry about 2 out of 5 figuring it out every second trial, which would be about the case if the lower bound would be the true value. Hence this seems to be a value that might indeed matter! The additional information may help the investigator to draw the appropriate conclusions from the experiment.

For the second experiment, we get a quite different result. Only two perceivers are found by means of our procedure. Hence we calculate $\hat{\delta}=\frac{2}{24}\approx 0.08$ for the lower bound, which is very small and quite similar to the estimator for the success probability above chance derived from the conventional approach of about 11%.

For the third experiment, using our approach we estimate the number of perceivers to be 11 out of 23. Hence we get $\hat{\delta}=\frac{11}{23}\approx 0.48$. In this case, the estimate for the success probability above chance derived from $\hat{\pi}$ in the conventional approach is given by about 30%. Again, we would claim that the results might lead to a different interpretation. Knowing that at least one half of the consumers will find the difference once in a while (but with a probability up to 60%) might be much worse than less than a third of the consumers finding the difference always. Maybe much smaller production costs could account for a loss of some of these purchasers, but most likely not up to a 50% loss.

Furthermore, the data set from the first example of Brockhoff and Schlich (1998) will be considered. The data can be found in table 2 and contains the results for $n = 12$ assessors and $k = 4$ replications each.

x	0	1	2	3	4	total
number of assessors	2	2	4	2	2	24

Table 2: Number of assessors with x right answers and total number of successes for the first experiment reported by Brockhoff and Schlich (1998).

In this experiment, using our approach we find only one perceiver out of 12 assessors, thus $\hat{\delta} \approx 0.08$. At the same time, the estimator for the success probability above chance is about 0.25. In this case, the lower bound can be assumed to be rather conservative. If the value of 0.25 is valid, we would conclude that at least three out of twelve assessors should have found the difference. This example shows that the true number of perceivers might indeed be reasonably larger than the lower bound derived from our approach.

Finally, we refer to the data presented by Priso, Danzart and Hossenlopp (1994). They consider $n = 6$ and $k = 10$ for the first respectively $n = 8$ and $k = 12$ for the second experiment. The outcomes are given in table 3.

x	0	1	2	3	4	5	6	7	8	9	10	11	12	total
exp. 1	0	0	2	1	1	0	0	2	0	0	0			25
exp. 2	0	0	2	1	2	0	2	0	0	0	1	0	0	37

Table 3: Number of assessors with x right answers and total number of successes for the two experiments reported by Priso *et al.* (1994).

For the first experiment, the estimated success probability above chance is about 13%, while we cannot even prove significant differences between these products, i. e. the 5% lower bound from our approach is zero. From this data, it might seem that there has to be at least one perceiver, otherwise we could not have observed two assessors with seven correct answers each. However, the p-value of the global test for product equality is 0.11. Using the method the other way round, for an effect size of 50%, i. e. an average success probability of $\frac{2}{3}$ (cf. Schlich 1993), we can show that at least one half of the assessors does not perceive the differences with such a large probability. Hence if we were only interested in that large effects, we could conclude that at least one half of the consumers will not find the difference that often.

For the second experiment, the conventional approach results in an estimate of about 8% for the success probability above chance, while once more we cannot even prove significant differences. Nevertheless, in this experiment it seems that there is again at least one perceiver: 10 successes out of 12 trials is very unlikely due to chance only, but here the number of assessors is too small to prove the alternative, as it is in the first example. Considering effect sizes according to Schlich (1993) for this example and considering the number of non-perceivers once again, for a 25% effect size we would have claimed that at least one assessor is a non-perceiver, while for a 50% effect size it would have been three of them.

Looking at the original data, we might have found one specialist with respect to this product among our assessors, whilst the greatest part of consumers does not really find any differences. This might explain why there is only one assessor with such a large number of successes, while the second-best assessor has only six. Hence a much more reasonable design might have been to use 24 assessors and let them replicate only four-times each, say.

Conclusions and outlook

The main concern of this paper was to propose an approach in order to estimate the minimal proportion of perceivers in repeated difference tests. By means of a simple example, we have shown that identical average success probabilities might be due to very different assessor performances. Here, the information about the minimal number of perceivers might be of great interest for the investigator. On the basis of a set of hypotheses that is closed unter intersection, we have developed a testing procedure. This procedure allows to test subsequent hypotheses while the type I error rate will not exceed the overall significance level α. With the help of multiple test theory we may therefore subsequently test whether there are at least $n-l$ non-perceivers out of n assessors, while l increases from 1 to n. If such a hypothesis is rejected, we know that there are at least l perceivers within the panel, i. e. assessors that might detect the difference between the products at least once in a while.

In several examples, we have shown that the proposed approach leads to reasonable lower bounds according to the given data structure. Thus we have an additional information to judge whether all successes have been given by a few assessors only or whether all assessors might have provided a similar number of successes. Contrariwise, also examples exist in which the lower bound will not provide any additional information compared to the average success probability of the assessors.

Resulting in a confidence bound, the outcomes of our method obviously depend on the chosen level α. As it has been shown in the examples, with a small number of assessors we sometimes may not identify any perceivers at all at a 95%-level, even though it seems that there have to be some. For a fixed total number of assessments we therefore propose the use of less replications in favor of more assessors whenever it is possible (cf. Meyners and Brockhoff, 2002). Furthermore, it has been stated that the method is rather conservative due to its construction.

If concerns are given to similarity testing, this approach may be easily adapted to derive an upper bound for the number of perceivers, while we may respect for the fact that only effect sizes of at least some pre-determined amount are of interest.

Finally, it has been stated that a distribution for the success probability of the perceivers could be assumed. For this, e. g., the Beta-binomial model might be used (Ennis and Bi 1998, Bi and Ennis 1999a, b). This model has been adapted by Brockhoff (2002) in order to save the success probability from falling below π_0, the success probability of pure guessing. The use of appropriate models may lead to more sophisticated estimates of the proportion under consideration. Maximum likelihood estimators may be derived in these cases. These have quite nice properties, but are relatively hard to determine and therefore rarely used. Furthermore, a textual justification of the assumptions is hard to determine, neither exist appropriate tests for these assumptions to the author's knowledge. Even more, with the approach presented here, misspecifications of the model might be detected whenever the estimated proportion of perceivers within the model falls below the bound derived by our approach. Hence this method might easily provide the investigator with some additional information. Even though the theory of our approach needs some technical efforts, the application of this method is considerably simple.

Acknowledgements

The author is grateful to the Deutsche Forschungsgemeinschaft (SFB 475, "Reduction of complexity for multivariate data structures") for the financial support of this work.

References

Bi, J. and Ennis, D.M. (1999a) *Beta-binomial tables for replicated difference and preference tests.* Journal of Sensory Studies 14, 347-368.

Bi, J. and Ennis, D.M. (1999b) *The power of sensory discrimination methods used in replicated difference and preference tests.* Journal of Sensory Studies 14, 289-302.

Brockhoff, P.B. (2002) *The statistical power of replications in difference tests.* to appear in Food Quality and Preference.

Brockhoff, P.B. and Schlich, P. (1998) *Handling replications in discrimination tests.* Food Quality and Preference 9, 303-312.

Ennis, D.M. and Bi, J. (1998) *The Beta-binomial model: accounting for inter-trial variation in replicated difference and preference tests.* Journal of Sensory Studies 13, 389-412.

Hunter, E.A., Piggott, J.R. and Lee, M.K.Y. (2000) *Analysis of discrimination tests.* Agro-industrie et methodes statistiques, Pau, january 19-21, 2000.

Kunert, J. and Meyners, M. (1999) *On the triangle test with replications.* Food Quality and Preference 10, 477-482.

Marcus, R., Peritz, E. and Gabriel, K.R. (1976) *On closed testing procedures with special reference to ordered analysis of variance.* Biometrika 63, 655-660

Meyners, M. and Brockhoff, P.B. (2002) *The design of replicated difference tests.* submitted for publication.

Meyners, M. (2002) *On the number of perceivers in a triangle test with replications.* In: C. Duby and J.P. Cassar (eds.): Actes des 7mes Journes Europennes Agro-Industrie et Mthodes Statistiques, Lille, january 16-18, 2002, 85-89.

Priso, H.E., Danzart, M. and Hossenlopp, J. (1994) *A statistical analysis of difference tests with replications.* Journal of Sensory Studies 9, 121-130.

Schlich, P. (1993) *Risk tables for discrimination tests.* Food Quality and Preference 4, 141-151.