**On repeated difference testing**

Joachim Kunert

## Abstract

If the number of assessors in a difference test is not large enough to ensure the desired power of the testing procedure, then it is often advised to use assessors repeatedly. That is, each assessor performs the testing not just once but several times. There is a discussion going on, how results of a repeated difference testing are to be analysed. The present paper (as was to be expected) supports the point of view expressed in Kunert and Meyners (1999). It also tries to generalise their approach such that we get confidence limits. While the exposition concentrates on the triangle test, the approach is also applicable for other difference testing procedures (e.g. pairwise difference test, duo-trio test).

Author's mailing address:

J. Kunert

Fachbereich Statistik

Universität Dortmund

D 44221 Dortmund

Germany

E-mail address: kunert@statistik.uni-dortmund.de

## 1. Introduction

In repeated difference tests, each assessor performs the testing procedure repeatedly, thus increasing the number of assessments. This introduces technical problems for the analysis of the experiment. How do we take account of the structure introduced by the repetitions? And, more fundamentally, what is the structure introduced by the repetitions?

A commonly used test statistic for repeated difference tests is the sum of all correct assessments, summed over all assessors. Several authors (e.g. o'Mahony, 1982, Brockhoff and Schlich, 1998) argue that the binomial distribution cannot be used to analyse this kind of data. Brockhoff and Schlich (1998) propose an alternative model for difference tests with replicates, where the assessors have different probabilities to correctly identify the odd sample even if the products are identical.

This is criticised by Kunert and Meyners (1999) who agree that assessors will have different probabilities of correct assessment if there are true differences, but who do not think that Brockhoff and Schlich's model makes sense under the null hypothesis of product equality. They show that, if the null hypothesis is true and the experiment is properly randomised and properly carried out, then all assessments are independent and all have the same success probability $c$. This implies that the sum of all correct assessments is binomial with parameter $c$, where $c$ depends on the special kind of experiment that we have done. For instance, we have $c$ = 1/3 for the triangle test, while $c$ = 1/2 for the duo-trio test. Therefore, the usual test based on this sum and the critical values of the binomial distribution is a level $\alpha$ test for the null hypothesis of equality of the products, even if there are replications.

If there are differences between the products, then things are more complicated. The present paper discusses the distribution of the sum of all correct assessments if the probabilities for correct assessment are not all equal to $c$. Let $\pi_{ij}$ be the probability that assessor $i$ gives a correct answer at his/her $j$-th assessment. To analyse the results of the experiment, we need a model for $\pi_{ij}$.

In the author's opinion any model for $\pi_{ij}$ must be compatible with the experimental setup. It has to be realised that the randomisation of an experiment restricts the possible distributions of the results. For instance, if there is no difference between the products, then $\pi_{ij}$ equals $c$, for each assessor $i$ and each assessment $j$, see Kunert and Meyners (1999). If there are differences between the products, then we assume that assessor $i$ with probability $a_i$ at his/her $j$-th assessment actually experiences the products, and not only guesses. Note that $a_i$ depends on the assessor but not on $j$, therefore we neglect possible variations of $a_i$ over time, e.g. due to fatigue.

With this assumption the probability to get a correct answer from assessor $i$ at any assessment is constant over time and equals

$$\pi_{ij} = a_i + (1 - a_i)c = c + (1 - c)a_i \ .$$

Here $c$ is the (constant) probability that we have had when all products were equal. Therefore, we can omit the index $j$ and define

$$\pi_i = c + (1 - c)a_i \tag{1}$$

as the probability of a correct answer from assessor $i$ at any given assessment. Let $Y_{ij}$ be the result of the $j$-th assessment by assessor $i$, where $Y_{ij} = 1$ if the assessment is correct and $Y_{ij} = 0$, otherwise. It is a central point of the paper to make clear that if the experiment is run properly, then for given $a_i$ the $Y_{ij}$ are conditionally independent. To achieve this independence, we must randomise the order in which the products are arranged for each presentation. And this randomisation must be done in such a way that it is independent of the other orderings and independent of the last assessment.

At any assessment, the assessor will either experience the difference and therefore give a correct answer, or he/she will only guess. We assume that the probability to experience the difference is a constant $a_i$ for each assessor. When the assessor does not experience the difference then he / she has to guess. With independent randomisation the assessor cannot influence his/her probability to guess correctly. It will always be $c$, independent of the outcome of the previous assessment. This is not the case if the randomisation is not done independently.

If e.g. we run a repeated triangle test with six replicates, then it is often recommended to give each assessor all six possible orderings AAB, ABA, BAA, ABB, BAB, BBA. It is only randomised which of the six orderings comes first, second, and so on, see e.g. Hunter (1996) or the ISO 4120 (1983).

Assume assessor $i$ in her first assessment experiences the difference and notices that the odd sample is in the middle, say. Further assume that in the second assessment she can not experience the difference. Then, however, she will expect that the odd product will not be at the same position as before. Therefore, she will take either the right or the left sample. With the ISO randomisation this gives the assessor a chance of 2/5 of a correct guess. Therefore, $Y_{i1}$ and $Y_{i2}$ are no longer independent. If $Y_{i1} = 1$ then the probability of $Y_{i2} = 1$ increases.

The assessor's strategy does not work, however, with independent randomisation for each assessment. Then, with whatever outcome of the other assessments, the assessor has probability $c$ to guess correctly if he/she does not experience the difference. Therefore with independent randomisation, the $Y_{ij}$ are independent, and for given $a_i$ we have that the number of correct answers $X_i$ from assessor $i$ is binomially distributed with success probability $\pi_i$.

To proceed, assume that our assessors are a sample from some super-population of potential assessors, such that the $a_i$ are i.i.d. random variables, with some unknown expectation $b$ and variance var($a_i$). Therefore, for given assessor $i$ and for all assessments $j$, the probability of correct assessment $\pi_i$ is a random variable $\pi_i = c + (1 - c)a_i$, with expectation $\pi^* = c + (1 - c)b$.

Note that the parameter $b$ is of interest: It is the probability that a randomly selected assessor correctly experiences the difference between the products at any given assessment, while $\pi^* = c + (1 - c)b$ is the probability of a correct answer at any given assessment with any randomly selected assessor.

We derive a confidence interval for the parameter $b$, using a bound for $\text{var}(a_i)$. The derivation also uses the central limit theorem, so the confidence interval is valid only if the number of assessors is not too small. We show some simulation results on the performance of the interval for small numbers of assessors.

Finally, the confidence interval is compared to corresponding intervals derived from other methods. This is done with data from an empirical study reported by Hunter, Piggott and Monica-Lee (2000).

## 2. A conservative confidence interval for *b*

We start with the conditional distribution of $X_i$ the number of correct answers of assessor $i$ for fixed $a_i$. Model (1) implies that, for fixed $a_i$, $X_i$ is binomially distributed with parameters $m$, where $m$ is the number of assessments performed by each assessor, and $\pi_i = c + (1 - c)a_i$. This implies that the conditional expectation is

$$\mathrm{E}(X_i \mid a_i) = m\pi_i = mc + m(1-c)a_i ,$$

while the conditional variance equals

$$\mathrm{Var}(X_i \mid a_i) = m\pi_i(1-\pi_i) = mc(1-c) + m(1-c)(1-2c)a_i - m(1-c)^2 a_i^2 .$$

Consequently, the unconditional expectation of $X_i$ is

$$\mathrm{E}(X_i) = mc + m(1-c)\,\mathrm{E}\,a_i = mc + m(1-c)b ,$$

while the unconditional variance equals

$$\mathrm{Var}\,X_i = \mathrm{E}(\mathrm{Var}(X_i \mid a_i)) + \mathrm{Var}(\mathrm{E}(X_i \mid a_i))$$

$$= mc(1-c) + m(1-c)(1-2c)b - m(1-c)^2\,\mathrm{E}\,a_i^2 + m^2(1-c)^2\,\mathrm{Var}(a_i)$$

$$= m(1-c)(c+b-2bc) - m(1-c)^2 b^2 - m(1-c)^2 \operatorname{Var}(a_i) + m^2(1-c)^2 \operatorname{Var}(a_i)$$

$$= m(1-c)(1-b)\big(c+(1-c)b\big) + m(m-1)(1-c)^2 \operatorname{Var}(a_i)\,.$$

Note that like the $a_i$, the $X_i$ are also i.i.d. variables. Therefore, we have from the central limit theorem that $Z = X_1 + ... + X_n$, the total number of correct assessments, is approximately normal, if the number $n$ of assessors is sufficiently large.

We can use this approximate normality to get a confidence interval for $b$. Basically, there are two ways to do this.

First method: Estimate Var $Z$ from the data.

Define $S^2 = \frac{1}{n-1}\sum (X_i - Z/n)^2$ . Then $nS^2$ is a consistent estimate for Var $Z$. Consequently,

$$T_1^{(b)}(Z) = \frac{Z - \mathrm{E}\, Z}{\sqrt{nS^2}}, \text{ say,}$$

is approximately $IN(0,1)$ distributed. Note that E $Z$ = $nm(c+(1-c)b)$, and therefore $T_1^{(b)}(Z)$ depends on $b$. We conclude that

$$U_1(Z) = \min\left\{\frac{Z - nmc + t_{1-\alpha,n-1}\sqrt{nS^2}}{nm(1-c)}, 1\right\}$$

is an approximate upper $1-\alpha$ confidence limit for $b$, while

$$L_1(Z) = \max\left\{\frac{Z - nmc - t_{1-\alpha,n-1}\sqrt{nS^2}}{nm(1-c)}, 0\right\}$$

is an approximate lower $1-\alpha$ confidence limit for $b$. Here, $t_{1-\alpha,\, n-1}$ is the critical value of the t-distribution with $n-1$ degrees of freedom. The interval $[L_1(Z), U_1(Z)]$ is an approximate level $1-\alpha$ confidence interval for $b$.

A disadvantage of $[L_1(Z), U_1(Z)]$ lies in the fact that it does not take account of the special structure of the variance of $Z$. Note that for $b = 0$ the nonnegative variable $a_i$ must be 0 with probability 1. It follows that for $b$ = 0 we have Var($a_i$) = 0 and, consequently, we know that $\mathrm{Var}(Z) = nmc(1-c)$. Therefore, we do not have to estimate it. Hence, for small $b$, the interval $[L_1(Z), U_1(Z)]$ does not use all available information and will therefore be generally too large. We will also see with the help of simulations that the normal approximation for $T_1^{(b)}(Z)$ does not work too well for some distributions of $a_i$. Therefore, we use the

Second method: Use an upper bound for Var $Z$.

Note that $a_i \in [0,1]$. Therefore, we have that $a_i^2 \le a_i$ with probability 1, and $\mathrm{E}\, a_i^2 \le \mathrm{E}\, a_i = b$. This implies that

$$\mathrm{Var}\ a_i \le b - b^2$$

and consequently

$$\begin{aligned} \mathrm{Var}\, Z &\le nm(1-c)(1-b)\big(c + (1-c)b\big) + nm(m-1)(1-c)^2 b(1-b) \\ &= nm(1-c)c(1-b) + nm^2(1-c)^2 b(1-b)\,. \end{aligned} \tag{2}$$

Equality holds if the $a_i$ are Bernoulli-variables, i.e. if they are only 0 or 1.Equation (2) and the central limit theorem imply that

$$T_2^{(b)}(Z) = \frac{Z - nmc - nm(1-c)b}{\sqrt{nm(1-c)c(1-b) + nm^2(1-c)^2 b(1-b)}}\text{, say,}$$

is asymptotically $IN(0, \delta^2)$ distributed, where $\delta \le 1$. It follows that, for large $n$, the inequality

$$T_2^{(b)}(Z) \le u_{1-\alpha} \tag{3}$$

is true with a probability of at least 1 – $\alpha$. Note that $u_{1-a}$ is the critical value of the normal distribution and not of the $t$-distribution. The denominator of $T_2^{(b)}(Z)$ is a constant, not an estimate.

Once we have observed $Z$, we therefore can calculate a one-sided confidence interval by determining the set of all $b \in [0,1]$, such that equation (3) holds. If $Z \ge nmc$, then $T_2^{(b)}(Z)$ is decreasing and continuous in $b$ for all $b \in [0,1]$. Therefore, there is a number $L_2(Z)$, such that $T_2^{(b)}(Z) \le u_{1-\alpha}$ for all $L_2(Z) \le b \le 1$. If $Z < nmc$, then equation (3) is true for all $b \in [0,1]$ and we define $L_2(Z) = 0$.

With these definitions $L_2(Z)$ is conservative approximate lower confidence limit for $b$. Similarly to equation (3), we have that for sufficiently large $n$ the inequality

$$T_2^{(b)}(Z) \ge -u_{1-\alpha} \tag{4}$$

is true with a probability of at least $1 - \alpha$. Unfortunately, if $Z < nmc$ then $T_2^{(b)}(Z)$ is increasing in $b$ for small $b$ and decreasing for larger $b$. Therefore, it is possible that there are up to 2 values $b$ with $T_2^{(b)}(Z) = -u_{1-\alpha}$. This may lead to the counter-inductive result that equation (4) may not hold for very small and for large $b$, while it holds for intermediate $b$. In the special instance $n = 1$, $m = 12$ and $c = 1/3$ and $Z = 0$, then $T_2^{(b)}(Z) = -1.96$ for $b = 0.036$ and for $b = 0.52$. Therefore, all $b < 0.036$ and all $b > 0.52$ would be rejected as too large, while all $b$ between 0.036 and 0.52 would be accepted.

To get an upper confidence limit, we define $U_2(Z) = 1$, if $T_2^{(b)}(Z) \ge -u_{1-\alpha}$ for all $b$, and we define $U_2(Z) = 0$ if $T_2^{(b)}(Z) < -u_{1-\alpha}$ for all $b$. If there are one or two $b_0$ such that $T_2^{(b_0)}(Z) = -u_{1-\alpha}$, then we take the conservative approach and define $U_2(Z)$ as the largest of those. This makes $U_2(Z)$ an approximate and conservative upper confidence limit for $b$. The conservative approach accepts some $b$ for which inequality (4) does not hold, i.e. it is too long. However this is not a real problem. It can be shown that the set of all Z for which there are two $b_0$ with $T_2^{(b_0)}(Z) = -u_{1-\alpha}$ has probability less than $\alpha/2$ whatever may be the true $b$.

Therefore, in what follows, we will use $[L_1(Z), U_1(Z)]$ as a $1 - 2\alpha$ confidence interval for $b$. It is easiest to calculate $L_2(Z)$ and $U_2(Z)$ numerically from equations (3) and (4) with the help of a computer.

## 3. Simulations on the performance of the interval

For $n = 1$ and for some distributions of $a_1$, the normal approximation can be a very poor fit for the distribution of $Z$. The approximation is relatively good in the case that $a_1$ has zero variance, i.e. if $a_1$ is always equal to $b$. In that case, $Z$ is binomially distributed, a distribution that generally can rather well be approximated by the normal distribution, especially if the number of replicates $m$ is large.

In the other extreme case, if $a_1$ is 1 with probability $b$ and 0 with probability $1 - b$, the distribution of $Z$ is very skewed. This implies that the normal distribution with the same mean and variance is no good fit.

With any distribution of the $a_i$ the fit gets better if $n$ gets large. It is of interest, therefore, to find out how well is the fit for reasonably sized $n$. To see how well our confidence limits work for $n$ in the practical range, we use simulations. The simulations are for $m = 12$ and for $n = 5$, 10 or 20. To see what happens in the worst case, we have simulated the most skewed situation in which each $a_i$ has probability $b$ to become 1 and probability $1 - b$ to be 0. Note that this two-point distribution is exactly the case in which the upper bound for the variance used for the calculation of $T_2^{(b)}(Z)$ is exact. Therefore, if $[L_2, U_2]$ has a good coverage probability for the simulated distribution of the $a_i$, it will have an even higher one for other distributions with a smaller variance.

The simulations indicate that for $n \geq 10$ the normal approximation works rather well for $T_2^{(b)}(Z)$ but surprisingly poor for $T_1^{(b)}(Z)$. The poor performance of $T_1^{(b)}(Z)$ is due to the fact that there is a correlation between $Z$ and $S$.

For $r = 1$ or 2, whenever $T_r^{(b)}(Z)$ is larger than the critical value or less than minus the critical value, then the confidence interval $[L_r(Z), U_r(Z)]$ does not cover $b$. Therefore, if the confidence intervals were exact, then we would expect that all the entries in Table 1 were equal to 250. If we observe a lower number than 250 then the estimated coverage probability is higher than expected. Any entry that is much larger than 250 indicates that the confidence interval is not reliable.

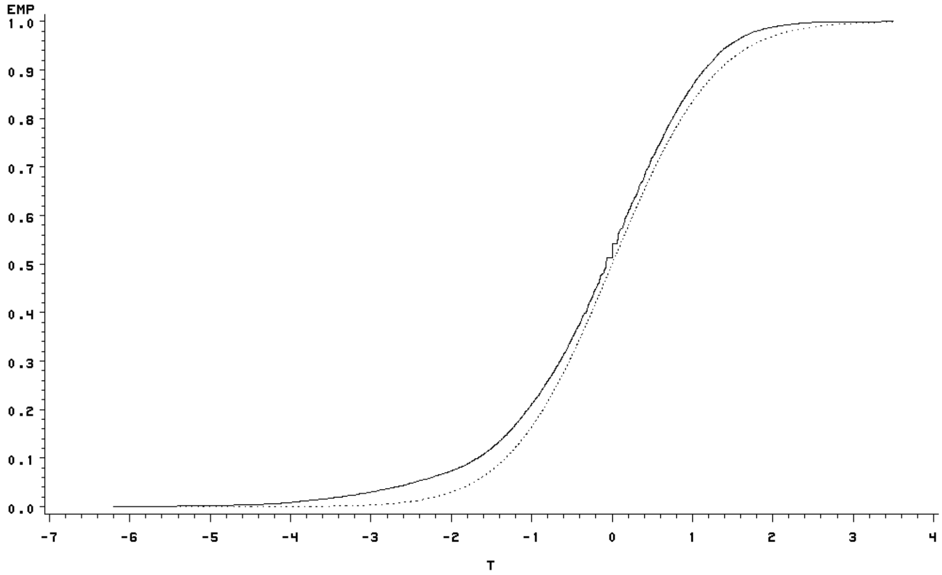
Table 1
Simulation results on the confidence intervals derived in section 2 when $m$ = 12

| $b$ | $n$ | number of runs (out of 10,000) where $T_1^{(b)}(Z) > t_{0.975,n-1}$ | $T_1^{(b)}(Z) < -t_{0.975,n-1}$ | $T_2^{(b)}(Z) > u_{0.975}$ | $T_2^{(b)}(Z) < -u_{0.975}$ |
|---|---|---|---|---|---|
| 0.0 | 20 | 249 | 279 | 265 | 221 |
| | 10 | 212 | 275 | 198 | 205 |
| | 5 | 200 | 303 | 201 | 174 |
| 0.1 | 20 | 72 | 722 | 279 | 139 |
| | 10 | 37 | 990 | 352 | 110 |
| | 5 | 34 | 762 | 499 | 47 |
| 0.2 | 20 | 112 | 516 | 300 | 200 |
| | 10 | 75 | 823 | 341 | 154 |
| | 5 | 56 | 1327 | 355 | 53 |
| 0.3 | 20 | 163 | 415 | 268 | 238 |
| | 10 | 138 | 379 | 273 | 165 |
| | 5 | 144 | 1096 | 339 | 83 |
| 0.4 | 20 | 258 | 303 | 273 | 243 |
| | 10 | 222 | 336 | 240 | 213 |
| | 5 | 252 | 678 | 171 | 195 |
| 0.5 | 20 | 299 | 215 | 229 | 262 |
| | 10 | 322 | 226 | 230 | 280 |
| | 5 | 449 | 284 | 335 | 224 |

The numbers for $T_2^{(b)}(Z)$ if $n \geq 10$ are not too far away from the expected 250.

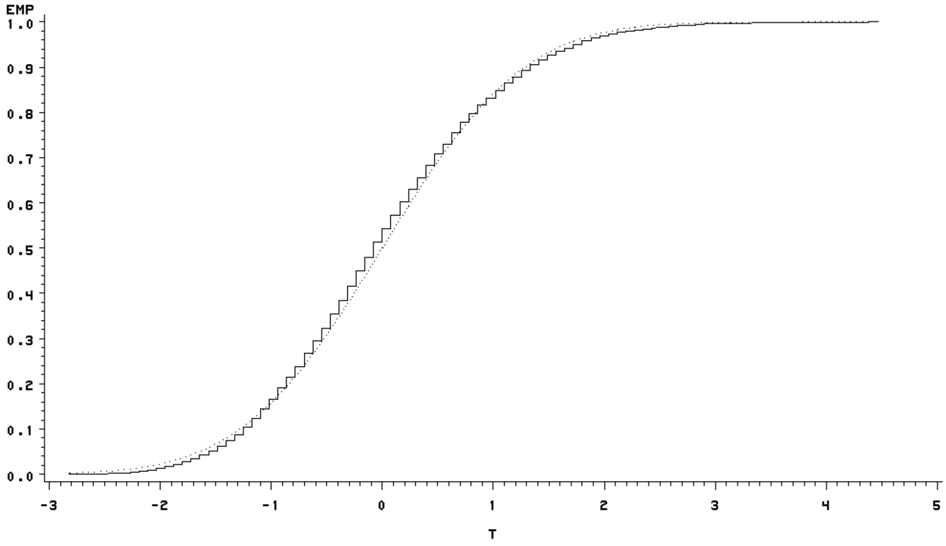
The statistic $T_1^{(b)}(Z)$, however, performs very poorly for instance for $b$ = 0.1, when we have too many cases where $T_1^{(b)}(Z) < -t_{0.975,n-1}$. Note that this means that we are falsely assuming that $b$ is significantly less than its true value. Figure 1 shows the empirical distribution function of $T_1^{(b)}(Z)$ in the case $m$ =12, $n$ = 20, $b$ = 0.1 and if the distribution of the $a_i$ is a two-point distribution. For comparison, Figure 1 also contains the distribution function of the t-distribution. It is evident that $T_1^{(b)}(Z)$ is consistently too small. This indicates why, even for $n$ = 20, we falsely conclude in 7% of the runs instead of the nominal 2.5% that $b$ is significantly less than 0.1.

Figure 1:
The empirical distribution function of $T_1^{(b)}(Z)$ in the case $m$ =12, $n$ = 20, $b$ = 0.1




For comparison, the empirical distribution function of $T_2^{(b)}(Z)$ displayed in Figure 2 fits much better to the normal distribution. However, here the distribution is slightly biased in the other direction: there are some observations less with small $T_2^{(b)}(Z)$ than expected from the normal distribution, while there are some observations with large $T_2^{(b)}(Z)$ too many.

Figure 2:
The empirical distribution function of $T_2^{(b)}(Z)$ in the case $m$ =12, $n$ = 20, $b$ = 0.1.



The poor performance of $T_1^{(b)}(Z)$ is due to the fact that there is a correlation between $Z$ and the estimated variance: Note that in our simulations each assessor is either a responder and therefore gives a correct answer all the time, or the assessor is a non-responder who always only guesses. For the data in Figure 1, the probability to be responder is 0.1. This implies that whenever there are many responders, then there are still many non-responders and the variance is large. If, however, there are no responders, then the variance is small. Therefore, in cases with large $Z$, we generally also had a large $S$ and $T_1^{(b)}(Z)$ was relatively small compared to $T_2^{(b)}(Z)$. In cases with small $Z$, we generally got a highly negative $T_1^{(b)}(Z)$ because $S$ was small. Therefore, the confidence interval based on $T_1^{(b)}(Z)$ is not reliable in the worst case scenario simulated here.

Our simulations seem to indicate that the confidence interval $[L_2(Z), U_2(Z)]$ is reliable for $n$ = 10 or more. It is of interest to see if it is useful. That is, we want to find out whether it is small enough not to cover too many $b$. It has to be noted that $[L_2(Z), U_2(Z)]$ was constructed to be model robust and conservative. Therefore, it will generally cover too many $b$. However, it was also constructed to be good for small $b$. This results in an excellent performance for testing the hypothesis that $b = 0$, i.e. that all products are equal.

We reject the null hypothesis of equality of the products whenever $[L_2(Z), U_2(Z)]$ does not cover 0. This, however, is equivalent to $T_2^{(0)}(Z) > u_{0.975}$. Note that $T_2^{(0)}(Z)$ is the test statistic which assumes that we have $n$ times $m$ assessors. Kunert and Meyners (1999) have shown that this is a level $\alpha$ test. The fact that for $b = 0$, we have $n\ m$ independent observations instead of $n$ makes the asymptotics work much better for this case.

Again simulating the two point distribution for the $a_i$, we have counted how often the null hypothesis of equality of the products is rejected by the test based on $[L_2(Z), U_2(Z)]$. Table 2 reports the results. For comparison, we have also counted how often $[L_1(Z), U_1(Z)]$ did not cover 0. It is clear from Table 2 that the confidence interval based on $T_2^{(b)}(Z)$ has much larger power.

Table 2
Number of experiments (out of 10,000) where the null-hypothesis of equality of the products is rejected for the two point distribution of the $a_i$.

| $b$ | $n$ | $T_1^{(0)}(Z) > t_{0.975,n-1}$ | $T_2^{(0)}(Z) > u_{0.975}$ |
|---|---|---|---|
| 0.0 | 20 | 249 | 265 |
| | 10 | 212 | 198 |
| | 5 | 200 | 201 |
| 0.1 | 20 | 1394 | 5215 |
| | 10 | 413 | 3715 |
| | 5 | 174 | 2761 |
| 0.2 | 20 | 4493 | 8682 |
| | | 1425 | 6729 |
| | | 315 | 4988 |
| 0.3 | 20 | 7549 | 9734 |
| | 10 | 3237 | 8618 |
| | 5 | 801 | 6809 |

If we simulate the one-point distribution, where all $a_i$ are equal to $b$ with probability 1, then we still find that the test based on $T_2^{(b)}(Z)$ gives more correct rejections for each $b > 0$ than the test based on $T_1^{(b)}(Z)$, see Table 3. The entries in Tables 2 and 3 for $b = 0$ indicate that both tests are level $\alpha$ tests. Note that, to get a 95% confidence interval, the one sided tests have to be level 2.5% tests. Therefore, we should have 250 rejections for $b = 0.0$.

We have simulated the two extreme situations for the distributions of the $a_i$. The two-point distribution is the one with the largest variance. The test based on $T_2^{(b)}(Z)$ is constructed with this distribution in mind. Therefore, it is no surprise that this test produced much more rejections of the null-hypothesis for this case. The one-point distribution is the case with the smallest variance. Here, the test based on $T_1^{(b)}(Z)$ should compete much better, because the test based on $T_2^{(b)}(Z)$ overestimates the variance. Our simulations show, however, that the latter performs better even in this case.

Table 3
Number of experiments (out of 10,000) where the null-hypothesis of equality of the products is rejected for the one point distribution of the $a_i$.

| $b$ | $n$ | $T_1^{(0)}(Z) > t_{0.975,n-1}$ | $T_2^{(0)}(Z) > u_{0.975}$ |
|---|---|---|---|
| 0.0 | 20 | 242 | 252 |
| | 10 | 200 | 215 |
| | 5 | 206 | 225 |
| 0.1 | 20 | 5,134 | 5,751 |
| | 10 | 2,411 | 3,093 |
| | 5 | 1,165 | 1,789 |
| 0.2 | 20 | 9,767 | 9,879 |
| | 10 | 7,405 | 8,387 |
| | 5 | 3,382 | 5,428 |
| 0.3 | 20 | 10,000 | 10,000 |
| | 10 | 9,742 | 9,937 |
| | 5 | 6,341 | 8,778 |

## 4. Practical examples

Hunter, Piggott and Monica-Lee (2000) report three experiments with repeated triangle tests. In all three experiments, each assessor examined $m$ = 12 presentations of the two products. In experiment 1 they had $n$ = 30, in experiment 2 they had $n$ = 24 and in experiment 3 they had $n$ = 23 assessors. So all three experiments were in the range where we would expect the interval $[L_2(Z), U_2(Z)]$ to be reliable.

Table 4 contains the data from the experiments.

Table 4
Number of assessors $i$ for which the number of correct assessments $X_i$ equals $x$

| $x$ | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Experiment 1 | 0 | 0 | 1 | 2 | 3 | 7 | 8 | 6 | 2 | 1 | 0 | 0 | 0 |
| Experiment 2 | 1 | 0 | 1 | 5 | 5 | 3 | 3 | 3 | 1 | 2 | 0 | 0 | 0 |
| Experiment 3 | 0 | 0 | 2 | 1 | 1 | 4 | 3 | 6 | 3 | 1 | 0 | 1 | 1 |

We count that $Z$ equals 170 in Experiment 1, 117 in Experiment 2 and 147 in Experiment 3. Clearly, the examples indicate that the model of a two point distribution of the $a_i$ with mass only at the points 0 and 1 is not correct for these data. All assessors with $a_i$ = 1 would have

$X_i$ = 12. It has to be pointed out, however, that the interval $[L_2(Z), U_2(Z)]$ does not assume this model to hold. It only allows it as a possible case.

If we determine the set of all $b$ for which equations (3) and (4) are fulfilled, we get that $[L_2(Z), U_2(Z)]$ equals [0.0862, 0.3919] in Experiment 1, [0.0151, 0.3057] in Experiment 2 and [0.1382, 0.5112] in Experiment 3. Note that none of the confidence intervals covers 0, so we can reject the null-hypothesis that all products are equal in all three experiments.

Hunter, Piggott and Monica-Lee (2000) in their paper considered $\pi^*$, the expected probability to get a correct result. They used several methods from the literature to calculate confidence intervals for $\pi^*$. It is interesting to note that these intervals from the literature make different assumptions on the distribution of $\pi^*$, while our interval is conservative and works without any assumptions on the distribution of the $a_i$. So is clear that it will generally give a larger interval. It is of interest to see how much larger it is. Also note that the methods considered by Hunter, Piggott and Monica-Lee (2000) do not split $\pi_i$ up into $c + (1 - c)a_i$.

To compare our results these confidence intervals, we therefore have to transform the interval $[L_2(Z), U_2(Z)]$ to $[c+(1-c)L_2(Z), c+(1-c)U_2(Z)]$, which is a model robust confidence interval for $\pi^*$.

We then get a generalisation of Hunter, Piggott and Monica-Lee's (2000) Table 2. The intervals displayed in Table 5 show that generally the model robust confidence interval is indeed the largest. There is one important exception: The model robust confidence interval is the only one (except the one based on the binomial distribution which is clearly too optimistic) to find out that there is a significant difference between the products in Experiment 2. The other models do not find a significant difference for this example.

We might get a smaller (and maybe more useful) confidence interval by making more restrictive assumptions on the distributions of the $\pi_i$ which exclude our two point distribution, but which use the relation $\pi^* = c + (1 - c)b$. Some work in this area is currently being done.

Table 5
Confidence intervals for p* from several methods.

| Method | Experiment 1 | Experiment 2 | Experiment 3 |
|---|---|---|---|
| Binomial Distribution* (assuming a one point distribution of the $a_i$) | [0.4206,0.5238] | [0.3495, 0.4629] | [0.4737, 0.5915] |
| GLM* | no interval derived | [0.3305, 0.4807] | [0.4419, 0.6214] |
| Brockhoff & Schlich* | [0.4207, 0.5238] | [0.3176, 0.4949] | [0.4297, 0.6355] |
| t-test (calculated from $[L_1(Z), U_1(Z)]$) | [0.4230, 0.5214] | [0.3277, 0.4848] | [0.4438, 0.6214] |
| model robust (calculated from $[L_2(Z), U_2(Z)]$) | [0.3933, 0.5947] | [0.3433, 0.5373] | [0.4254, 0.6742] |

*) These intervals were calculated by Hunter, Piggott and Monica-Lee (2000).

**Acknowledgement:**

Financial support from the Deutsche Forschungsgemeinschaft (SFB 475, "Reduction of complexity in multivariate data structures") is gratefully acknowledged.

## References


BROCKHOFF, P. B. and SCHLICH, P. (1998): Handling replications in discrimination tests. *Food Quality and Preference* **9**, 303 – 312.

HUNTER, E. A. (1996): Experimental design. *In:* Næs, T. and Risvik, E. (ed.): *Multivariate analysis of data in sensory science*, Elsevier, Amsterdam, 37 - 69.

HUNTER, E. A., PIGGOTT; J.R. and MONICA-LEE, K.Y. (2000): Analysis of Discrimination Tests. *Agro-industrie et methodes statistiques, Pau* 19-20 *Janvier* 2000, Proceedings.

ISO 4120 (1983): Sensory Analysis - Methodology - Triangular Test.

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

O'MAHONY, M. (1982): Some assumptions and difficulties with common statistics for sensory analysis. *Food Technology* **36**, 75 – 82.