Available online at www.sciencedirect.com
SCIENCE DIRECT®
Food Quality and Preference 17 (2006) 209–218
Food Quality and Preference
www.elsevier.com/locate/foodqual

Comparison of ANOVA with the Tobit model for analysing sensory data

M. Marin-Galiano *, J. Kunert

Department of Statistics, University of Dortmund, D-44221 Dortmund, Germany

Received 25 September 2004; received in revised form 25 May 2005; accepted 29 July 2005
Available online 21 September 2005

Abstract

The data of sensory trials often contain a large number of zeroes, due to the limited scale used. It therefore is clear that the data are not normal. Out of concern that this might lead to problems with the application of ANOVA, Guillet et al. [Guillet, M., Methot, S., & Rodrigue, N. (2001). Application of Tobit models to handle zero-valued attribute intensities. *Presented at the Pangborn conference in Dijon*] proposed to use the Tobit model for the analysis of sensory trials. They demonstrated with a data set for a workshop at the fourth Pangborn Sensory Science Symposium that this model generally detects more significant differences between products than ANOVA does.

It should be noted, however, that randomization theory provides a justification to use ANOVA for designed experiments, even for non-normal data. On the other hand, the Tobit model has strict model assumptions itself, and the usual proof of the consistency of the maximum likelihood estimate in the Tobit model does not work for sensory trials.

Using the same data set as Guillet et al. [Guillet, M., Methot, S., & Rodrigue, N. (2001). Application of Tobit models to handle zero-valued attribute intensities. *Presented at the Pangborn conference in Dijon*], we compare the two models with the help of permutation tests. Our results indicate that ANOVA allows to test without violating the nominal level, while the Tobit model rejects the null hypothesis too often.

Keywords: Sensory analysis; ANOVA; Tobit model; Permutation tests; *F*-test; *t*-Test

1. Introduction

Analysis of variance (ANOVA) is a powerful and popular tool in statistical inference for the comparison of products. Its popularity, among other reasons, is due to the well-established fact that it is quite robust to non-normality of the data (cf. Conover, Johnson, & Johnson, 1981). In general, ANOVA tends to be less affected by violations of the model assumptions than other methods.

In this article, we deal with the situation in which trained assessors quantify the intensities of sensory variables on a variety of products by using a limited scale. Such a scale generally starts with a value of zero, meaning that there is no intensity, and ends with a chosen maximum number for a very high intensity. If some products contain a very low intensity for a given variable, this will cause even trained assessors to notice no intensity at all, resulting in a large amount of zeroes. An example for such data was used for a workshop at the Pangborn Sensory Science Symposion in Dijon in 2001. This data set, which we call the Pangborn data set in what follows, contains the assessments of 23 dairy products with strawberry flavor by 13 assessors, evaluating 24 sensory variables. The workshop, in which the

* Corresponding author. Tel.: +49 231 755 3113; fax: +49 231 755 3454.
E-mail addresses: marcos.marin-galiano@uni-dortmund.de (M. Marin-Galiano), kunert@statistik.uni-dortmund.de (J. Kunert).

0950-3293/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.foodqual.2005.07.016

data was presented, contrasted several approaches to analyse these data. The large amount of zeroes makes it clear that the assumption of normality will be violated in this case. But one of the participants, Meyners (2001), showed by using a permutation test that despite this violation, the usual F-test from ANOVA for equality of all products appears to keep the nominal level α, since the empirical distribution of the F-test statistics matches the appropriate F-distribution.

On the other hand, there was a contribution by Guillet, Methot, and Rodrigue (2001), who stress a number of problems of the data, namely non-normality and heteroscedasticity. Therefore, they conclude that ANOVA is not suitable for these data and propose to use the Tobit model instead. The Tobit model (cf. Tobin, 1958 or Amemiya, 1985) was developed in econometrics for analysing data of household expenses, which have a high amount of zero data for luxury goods. It combines regression analysis with a probit approach to get rid of the disadvantages arising from using one of these methods alone. Guillet et al. (2001) state that the Tobit model is a generalization of ANOVA: if there is a small number of zeroes in the data, the results of ANOVA and Tobit will be nearly the same. But if there is a large number of zeroes in the data, the estimated differences among the products will increase in the Tobit analysis. Hence, the Tobit model has a greater ability to detect differences. The question arises whether differences found by Tobit are reliable.

This article is structured as follows. In Section 2 we describe the Tobit model as presented by Tobin (1958) and Amemiya (1985). After giving a short summary of the historical background of the model, we outline the model with its assumptions. The Tobit model is not widely known, so we give a brief summary of how to derive the likelihood function of the model which is necessary to estimate the parameter vector. Hypothesis tests can then be performed by either using a likelihood ratio test or by using the asymptotic variance of the estimates. This asymptotic variance is derived by Amemiya (1985) if a number of assumptions are fulfilled. In Section 3 we give a short description of the Pangborn data set. We describe the experimental setting of the data and discuss whether all assumptions of our two models are true for this setting. In Section 4 we show the results of our analysis. Finally, we give a summary of our findings in Section 5.

2. The Tobit model

2.1. Background

The Tobit model was created by Tobin (1958) as a solution for problems with certain econometric data. This data contained the results of a survey comparing household incomes and expenses for a number of goods. For luxury goods, it is observed that households with a low income will spend no money on them at all. Therefore, trying to model the relationship between income and expenses for luxury goods by using one linear regression for both low and high incomes leads to a very poor fit. Furthermore, the regression line becomes negative for households with a very low income, which clearly makes no sense.

A possible alternative might be a probit analysis, simplifying the data to a zero–one variable, simply measuring whether the household has spent money for this luxury good or not. The drawback of this method is the loss of information, since it neglects the size of the observations which are not zero.

A local regression with two lines, one at zero for low incomes and one with an appropriate slope for high incomes, seems to be preferable. Therefore, Tobin's proposal is to combine both approaches. The next subsection outlines his model in detail, using the notation of Amemiya (1985).

2.2. The model

For n observations, the Tobit model consists of latent variables y_i^* and observed variables y_i. Formally, the latent variables y_i^*, $i = 1, 2, \ldots, n$, fulfill a linear model

$$y_i^* = x_i'\beta + u_i, \tag{1}$$

while the observations y_i, $i = 1, 2, \ldots, n$ are derived as

$$y_i = \begin{cases} y_i^*, & \text{if} \quad y_i^* > 0, \\ 0, & \text{if} \quad y_i^* \leqslant 0. \end{cases} \tag{2}$$

Here, the u_i are i.i.d. normally distributed random variables with expectation 0 and unknown variance σ^2, $\beta \in \mathbb{R}^k$ is the unknown vector of parameters that we want to estimate and $x_i \in \mathbb{R}^k$ is a known vector of regression variables for the ith observation.

It is visible that this model is a hybrid model, consisting of a regression in (1) and a threshold component in (2), very similar to a probit model.

Consider the matrix $X \in \mathbb{R}^{n\times k}$ whose ith row is x_i'. To derive asymptotic properties of the maximum likelihood estimate $\hat{\beta}$, Amemiya (1985) made the following assumptions:

(T1) The x_i are uniformly bounded.
(T2) The limit $\lim_{n\to\infty} n^{-1}X'X$ exists and is positive definite.
(T3) Assume $\theta = (\beta', \sigma^2)'$ and let $\theta_0 = (\beta_0', \sigma_0^2)'$ be the true value for θ. The parameter space Θ is then compact, does not contain the region with $\sigma^2 \leqslant 0$ but contains an open ϵ-neighbourhood around θ_0.

We will discuss these assumptions in detail in Section 3.3 and discuss the asymptotic properties later in this section. First we show a way how to calculate $\hat{\beta}$.

2.3. Fitting the model using maximum likelihood estimation

Since the model (1) and (2) contains a probit component, it seems sensible to use the maximum likelihood estimates for β and σ^2. From Eqs. (1) and (2) the likelihood function can be derived as follows. For given i we have

$$P(y_i = 0 \mid x_i, \beta, \sigma^2) = P(y_i^* \leqslant 0) = P(x_i'\beta + u_i \leqslant 0)$$
$$= P(u_i \leqslant -x_i'\beta) = P\left(\frac{u_i}{\sigma} \leqslant \frac{-x_i'\beta}{\sigma}\right) = \Phi\left(\frac{-x_i'\beta}{\sigma}\right) \quad (3)$$

and similarly for $a_i > 0$

$$P(y_i \leqslant a_i \mid x_i, \beta, \sigma^2) = P(u_i \leqslant a_i - x_i'\beta) = \Phi\left(\frac{a_i - x_i'\beta}{\sigma}\right). \quad (4)$$

Combining (3) and (4) we get the distribution function of y_i

$$F(y_i \mid x_i, \beta, \sigma^2) = \begin{cases} 0, & \text{if} \quad y_i < 0, \\ \Phi\left(\frac{-x_i'\beta}{\sigma}\right), & \text{if} \quad y_i = 0, \\ \Phi\left(\frac{y_i - x_i'\beta}{\sigma}\right), & \text{if} \quad y_i > 0. \end{cases}$$

The distribution of y_i therefore is a mixture between a discrete and a continuous distribution. All y_i^* below 0 are aggregated at 0, leading to a jump of the distribution function at 0. For all $y_i > 0$ the distribution function is continuous and we have the continuous density function

$$f(y_i \mid x_i, \beta, \sigma^2) = \frac{1}{\sigma}\phi\left(\frac{y_i - x_i'\beta}{\sigma}\right),$$

while at $y_i = 0$ we have the discrete density

$$f(0, x_i \mid \beta, \sigma^2) = \Phi\left(\frac{-x_i'\beta}{\sigma^2}\right) = 1 - \Phi\left(\frac{x_i'\beta}{\sigma^2}\right).$$

Here Φ is the distribution function and ϕ is the density of the standard normal distribution. Given a sample of y_i for $i = 1, \ldots, n$, the likelihood can therefore be written as

$$L(\beta, \sigma^2) = \prod_{i=1}^{n} f(y_i \mid x_i, \beta, \sigma)$$
$$= \prod_{y_i = 0}\left(1 - \Phi\left(\frac{x_i'\beta}{\sigma}\right)\right) \cdot \prod_{y_i > 0}\frac{1}{\sigma}\phi\left(\frac{y_i - x_i'\beta}{\sigma}\right). \quad (5)$$

We have used an iterative algorithm to find the maximum of (5), namely the Newton–Raphson-algorithm with step-halving and Fisher scoring. The first and second derivates of the logarithm of (5), which must be calculated to use Newton–Raphson, can be found in Amemiya (1985). There are some discussions on how to find an optimal starting point for the algorithm. Tobin (1958) himself used an estimator, which is based on an approximation of Mill's ratio. Amemiya (1973) shows that this estimator is inconsistent and proposes a consistent estimator based on quadratic regression. Fair (1977) proposes to simply use $\beta = 0$ and $\sigma = 1$ as starting point.

Note that the ML-estimates in the Tobit model can also be calculated using commercial software. For instance, SAS can fit the Tobit model using PROC LIFEREG.

2.4. Hypotheses tests

The question whether some independent variables have an impact on the dependent variable, can be answered by a likelihood ratio test. The test problem can be formulated as

$$H_0 : \beta_{i_1} = \cdots = \beta_{i_q} = 0 \quad \text{vs}$$
$$H_1 : \quad \text{at least two of the } \beta_{i_j} \text{ are not equal,}$$

given that $q \leqslant k$ and the set of numbers $\{i_1, \ldots, i_q\}$ is a subset of $\{1, \ldots, k\}$. Let L_{full} be the maximum value of the likelihood function from a full model including all β_i. Accordingly, let L_{rest} be the maximum value of the likelihood function from the restricted model, in which all β_i mentioned in H_0 are set to zero. Then the test statistic is

$$\lambda = -2\ln(L_{\text{full}} - L_{\text{rest}})$$

and the null hypothesis can be rejected at the level α if $\lambda > \chi^2_{q,1-\alpha}$, where $\chi^2_{q,\alpha}$ is the α-quantile of a χ^2-distribution with q degrees of freedom.

2.5. Properties of the maximum likelihood estimator

The following theorem contains some properties of the maximum likelihood estimator of the parameters in (1) and (2).

Theorem 1. *Assume model (1) and (2) holds. Let L be the likelihood function given in (5). Furthermore, let $\hat{\theta}_n$ be a solution maximising* $\ln L$. *If the Jacobian-matrix of* $\ln L$ *is not singular and the assumptions given in Section 2.2 are fulfilled, then the following properties of $\hat{\theta}_n$ are true*:

(1) *$\hat{\theta}_n$ is a strongly consistent estimator for the true θ_0.*
(2) *$\ln L$ has a unique maximum at $\hat{\theta}_n$.*
(3) *The asymptotic distribution of $\hat{\theta}_n$ is normal, more precisely*

$$\mathscr{L}(\sqrt{n}(\hat{\theta}_n - \theta_0)) \to_w N\left(0, \left[-\frac{1}{n}\frac{\partial^2 \ln L(\theta_0)}{\partial\theta\,\partial\theta'}\right]^{-1}\right).$$

The proof of property (1) and (2) can be found in Amemiya (1973), property (3) was proved by Olsen (1978). It follows that all starting points of an iterative procedure like the Newton–Raphson-algorithm will lead

to the same unique estimator. However, the results are only true if the assumptions of the Tobit model in Section 2.2 are fulfilled. In Section 3.3 we discuss whether this is the case for the Pangborn data. Note that part (3) of the theorem can be used to approximate the variance of estimates $\ell'\hat{\theta}$.

3. The Pangborn data set and its problems

3.1. The experimental setting

The Pangborn data set (provided by Danone in 2001) was derived from an experimental setting as follows. A total of 23 dairy products with strawberry flavor were presented to a panel of 13 trained assessors. The order of the products was randomized for each assessor. In the data set, the products are labelled by a four-digit codeword, consisting of letters and numbers. Each assessor evaluated the intensity of the products for 24 sensory attributes. This was performed by marking the experienced intensity on an unstructured scale from 0 (no intensity) to 9 (maximum intensity). The assessors performed up to two replications on all products. To simplify the analysis, we only considered the first replicate for our analysis, meaning that we could work in the simple block model with assessors as blocks.

As is usual in sensory profiling, the meaning of the attributes is not understandable for an outsider. Some attributes have an intuitively appealing name, e.g., CREAMY, CITRUS or JAM. The names of some other variables seem to be clear, but are split into sub-variables, e.g., MILK and MILK2, the difference between these two is not clear. Finally, there are variable names which have no intuitive appeal at all, e.g., GREEN and VEFR. However, this is not a problem for the purpose of our study. The assessors themselves were trained by giving some physical reference for all attributes.

3.2. Problems with the assumptions of ANOVA

For regression analysis, many textbooks recommend to analyse the residuals "to check whether the assumptions of regression analysis are met:

(i) the errors are independent,
(ii) the errors have zero mean,
(iii) the errors have a constant variance,
(iv) the errors follow a normal distribution",

see, e.g., Draper and Smith (1981, p. 141ff).

Due to the similarity between ANOVA and regression analysis, we might want to check the same four conditions for the Pangborn data set. We immediately get doubts about the normality of the data. As shown in Fig. 1, a certain percentage of the data is zero. This percentage varies among variables and only two attributes (FRUI.INT and MILK.INT) have nearly no zero data. In other attributes, the percentage of zero data can reach up to 40%. Taking account of the fact that the intensities were marked on an unstructured scale, it is possible that the markings for null values could get slightly out of place. If we count all values below 0.3 as "no intensity found", the percentages of zero values are even higher. Hence, it is clear that the data and, therefore, the errors are not normal.

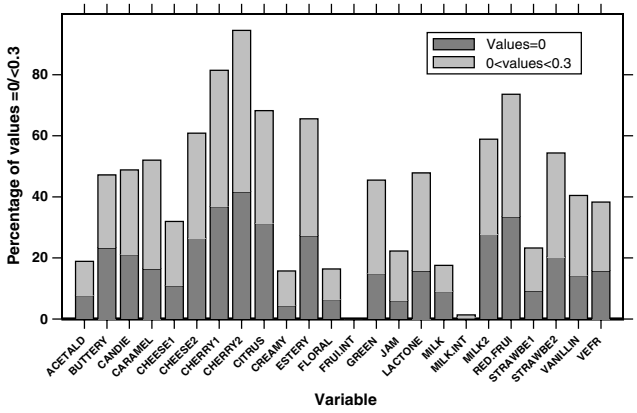

Fig. 1. Percentages of judgements equaling zero (dark shade) and judgements being lower than 0.3 (dark + light shade) for all variables.

To check the other three conditions, we might produce a plot like Fig. 2. The figure plots the residuals against the predicted value for the variable MILK2 in the simple block model with assessors as blocks. The diagonal in Fig. 2 is formed by the observations at 0. It can be seen that the residuals are clearly not independent of the predicted value. In regression analysis, such a residual plot would cause the statistician to not use this model.

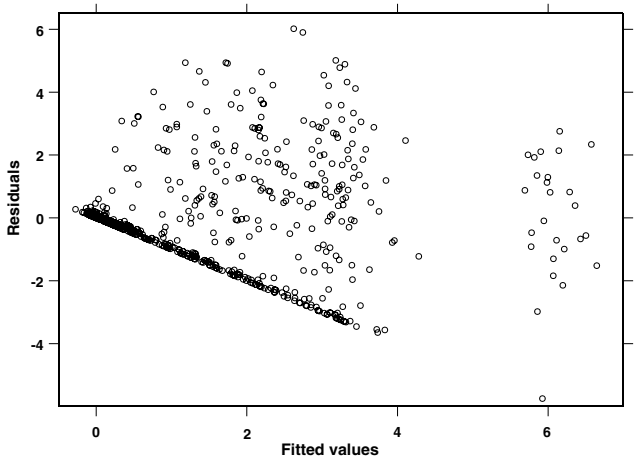

Fig. 2. Plot of residuals against predicted values of the attribute MILK2 for an ANOVA using assessor and product effects.

A final point stressed by Guillet et al. (2001) is the heteroscedasticity of the data. Using Hartley's test for homogenity of variance, Guillet et al. (2001) show that the test rejects the homogenity for nine attributes. Especially for attributes with a high amount of zeroes, the values of the test statistic can be enormous (the highest value of the test statistic is 2209 for variable CHERRY2, whereas the critical value of the test for a test level of $\alpha = 0.05$ is 4.8).

Note, however, that the data come from a designed experiment, in which the order of presentation of the products to the assessors was randomized. It then can be shown under a randomization theory viewpoint that the first three points required for regression analysis are not a problem. If we use a model that fits to the randomization used, this guarantees that the usual estimates of product differences from ANOVA are unbiased, no matter whether the data are normal or not. Furthermore, the estimate of the variance of such an estimate is also unbiased, see e.g., Bailey (1981) for details. This does not guarantee, however, that the test statistics derived from ANOVA have the χ^2-distribution or t-distribution that they should have under normality. There is some robustness to nonnormality, though, due to the central limit theorem. The estimates derived by ANOVA generally are the means over a number of observations. It therefore is no surprise that Kunert, Meyners, and Erdbrügge (2002) could show for sensory data that in most situations with non-normal data the distribution of the test statistics is very near to the required distribution. But they have also found instances, when the non-normality of the data gets too extreme. This may happen if a vast majority of the data are all the same number, while there are only a few outliers.

3.3. Problems with the assumptions of the Tobit model

In the preceding paragraph we found that there might be problems with the application of ANOVA on the Pangborn data set. We now check whether the assumptions of the Tobit model outlined in Section 2.2 are realistic for these data.

The strongest assumption certainly lies in Eq. (1) which assumes that the latent observations y_i^* should be normally distributed with constant variance. At first sight, already, this does not seem very realistic, if we consider the high proportion of observations between 0 and 0.3 in Fig. 1. We therefore decided to take the same approach as Guillet et al. (2001) and consider all such observations as 0.

The proof of Theorem 1 further needs assumptions (T1)–(T3). These assumptions could easily be satisfied if we drew a sample from a given population. But in the Pangborn data we have a given number of assessors and of products, and each combination of assessors and products appears exactly once. The only possibility to let the sample size n go to infinity would be to add more products or more assessors. In this case, however, the number of parameters would increase as well. If n tends to infinity, the dimension of $X'X$ therefore will also go to infinity. This implies that the limit in (T2) does not exist.

In all, the assumptions (T1)–(T3) cannot be used for our setting. But these standard assumptions for the Tobit model are needed in Amemiya (1985) to prove the asymptotic properties of the maximum likelihood estimator outlined in Section 2.5. Hence, we have no proof that the maximum likelihood estimator is consistent, unique and asymptotically normally distributed.

Actually, things are even worse than that. It is well-known, see e.g., Section 10.5.3 in Amemiya (1985), that even if the conditions (T1)–(T3) are fulfilled, the asymptotic distribution of the estimate works only if the assumptions of the Tobit model, namely the normality of the latent variable y^* in (1), holds. Therefore, the Tobit model depends more on the normality of the data than ANOVA does!

The last two subsections showed that both methods, ANOVA and the Tobit model, may have problems due to the specific data structure of sensory data. It therefore is not clear that the model assumptions which ensure reliable results are satisfied. Nevertheless, it is possible that the violation of the model assumptions does not impact the reliability of the results, as outlined by Conover et al. (1981) or Meyners (2001) for ANOVA. A possibility to determine the reliability is the application of permutation tests.

4. Simulation study

4.1. The permutation strategy

To see whether significant differences between products derived with a statistical test are reliable or not, we might check how often this test falsely declares differences, if in reality there are none. One method to estimate this false discovery rate for real situations would be to apply the test on a large number of uniformity trials, that is on data derived from experiments where all products were in fact identical. There are two problems connected with this approach. Firstly, at least in the case of a sensory experiment, a uniformity trial would not produce realistic data. The assessors would realize that all products are the same and, therefore, they would behave differently. A way out of this problem is to mimic uniformity data by taking the response from an experiment with non-identical products and to permute the data for each assessor. For this permuted data, any differences between the products are due to chance. If we calculate the test statistic for the permuted data set, any significant result therefore is an observation of a false discovery.

A second problem with uniformity data is that, even in cases where uniformity trials are possible, we can never produce a large number of uniformity trials—this would be too costly. However, by permuting the same data set repeatedly, we can produce a large number of uniformity data sets. The empirical distribution function of the test statistics derived from this large number of permuted data sets is an estimate of the true distribution function of the statistics under the null hypothesis.

Here, we use this permutation strategy and the Pangborn data set to check whether the theoretical distributions of the test statistics for ANOVA and for the Tobit model match the empirical distributions derived from the permuted data. If the distributions are the same, we can conclude that the test statistic is well suited for analysing the data. Otherwise, counting the number of permuted data sets resulting in a significant test statistic enables us to compute an empirical test level for a given test. From this empirical test level we can conclude whether this test is conservative or not.

4.2. Simulation

All simulations were performed using the software packages S-PLUS 2000 and R (2004). To avoid problems with random or fixed subjects, we decided to use only the first replicate from each assessor, resulting in 299 judgements of 13 assessors for 23 products. The data set was randomized according to the experimental setting, meaning that all scores given by one assessor in one variable were permuted between the products. The scores of different assessors and for different variables were not permuted. To get an impression how a varying percentage of zeroes affects the empirical distribution of the examined test statistics, the permutation study was performed for all 24 variables in the data set. For each attribute, 4000 permutations were performed as described above.

For ANOVA, we then assumed the simple block model with assessors as blocks. That is, we assumed that the rating of a given attribute of the jth product by the ith assessor could be written as

$$y_{ij} = \tau_j + \alpha_i + u_{ij},$$

where the $u_{i,j}$ are i.i.d. errors, τ_j is the effect of product j and α_i is the effect of assessor i.

For the Tobit model, we first changed all observations less than 0.3–0, similar to the way the Tobit analysis was performed by Guillet et al. (2001). This was done to improve the plausibility of Eq. (1), assuming that maybe a marking in that narrow area was indeed meant to be 0. We then specified Eq. (1) in the same way as for ANOVA, i.e.,

$$y^*_{ij} = \tau_j + \alpha_i + u_{ij}.$$

For each of the permutations, we computed a number of corresponding test statistics of interest. Taking all these simulated statistics, empirical distribution functions of the test statistics can be calculated. Each empirical function can be compared to the theoretical distribution function that the corresponding test statistic should have. Furthermore, an empirical test level can be obtained by calculating the relative frequency of simulated test statistics which are in the rejection area of this test, for some given nominal level α.

4.3. Results

4.3.1. Tests for equality of all products

In the case that we want to test whether all products have the same intensity for one of the attributes, we use a standard F-test for ANOVA and the likelihood ratio test as outlined in Section 2.4 for the Tobit model.

The closed lines in Figs. 3 and 4 show the empirical distribution function of the test statistics that we obtained from the permuted data sets. We also included

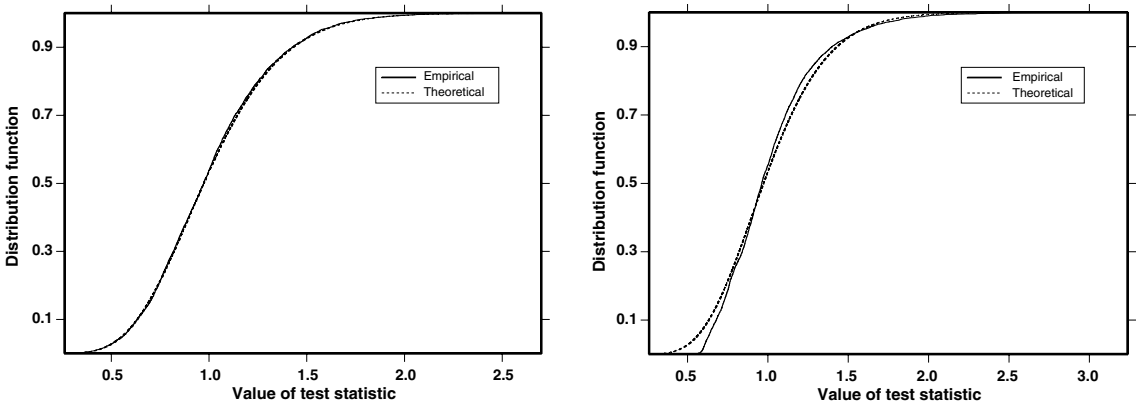

Fig. 3. Comparison of the empirical and theoretical distribution function for the test for equality of all products when using ANOVA. The theoretical distribution of the F-test statistic is the F-distribution with 22 and 264 degrees of freedom. Left: best fit for variable STRAWBE1. Right: worst fit for variable CHERRY2.

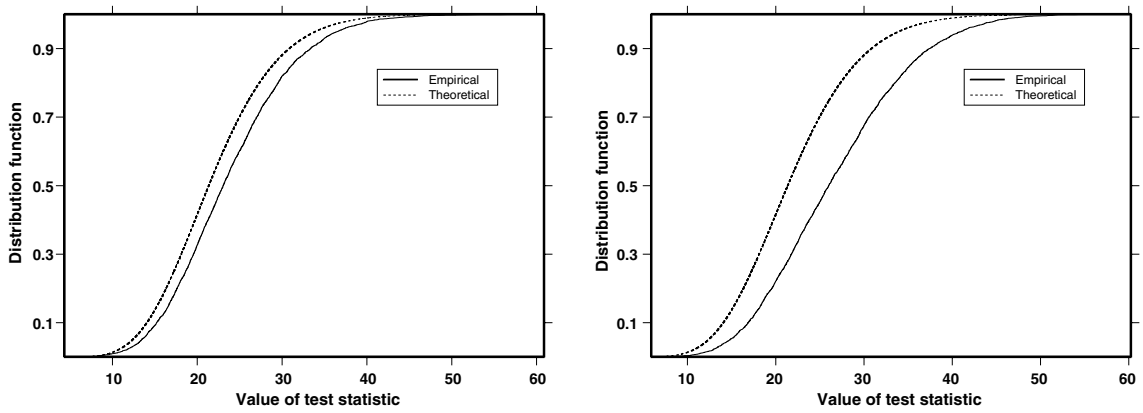

Fig. 4. Comparison of the empirical and theoretical distribution function for the test for equality of all products when using the Tobit model. The theoretical distribution of the χ^2-test statistic is the χ^2-distribution with 22 degrees of freedom. Left: best fit for variable STRAWBE1. Right: worst fit for variable CHERRY1.

a dotted line with the distribution function that we would expect in theory. These theoretical distributions are the F-distribution with 22 and 264 degrees of freedom in the case of ANOVA and the χ^2-distribution with 22 degrees of freedom in the case of the Tobit model. The figures display the best and worst fit that can be found among all 24 variables.

Fig. 3 gives an impression of the robustness of ANOVA. The best fit, measured by the average horizontal distance between the two curves, can be found for the variable STRAWBE1. As shown on the left hand side of Fig. 3, the theoretical and empirical distribution function agree very well, so that there is no visible difference between the curves. For most of the other variables, the comparison of the two distribution functions results in a very similar picture. Some slight differences are visible only for the variables CHERRY1 and CHERRY2. The worst fit can be seen in the variable CHERRY2, which is the variable with the highest percentage of zero data in the data set (cf. Fig. 1). Note, however, that when performing an F-test, we are only interested in the quantiles of 0.9 or higher (based on your level α). In this region, the match between the curves seems to be very well. The main differences occur for quantiles below 0.25 or between 0.7 and 0.9. It seems that the F-test might be a good choice, although some conditions of ANOVA are not fulfilled.

The same analysis for the Tobit model shows a completely different situation. As shown in Fig. 4, the empirical distribution is clearly shifted to the right. This gap is visible even in the best case which is the variable STRAWBE1 and grows with increasing percentage of zeroes, until the worst case which is the attribute CHERRY1. It implies that the critical value of the χ^2-test is reached too often and is an obvious indication for an anti-conservative test.

In order to strengthen these results, we counted the number of test statistics exceeding the respective critical value, using the theoretical levels $\alpha = 0.01$, $\alpha = 0.05$ and $\alpha = 0.1$. Since this number is binomially distributed with parameters $n = 4000$ and π, where π is the true test level, we can estimate the true test level and derive a 95% confidence interval for it. The results are displayed in Fig. 5. They confirm the results obtained by the comparison of theoretical and empirical distribution function in Figs. 3 and 4. When applying ANOVA, the theoretical test level lies in the 95% confidence interval for the true level, for all variables except CHERRY2. For the Tobit model, however, the empirical levels are twice to six times as large as the respective theoretical levels. There is only one exception of this rule, namely the variable CHERRY2 for a theoretical test level of $\alpha = 0.01$. In this case only, the Tobit model performs better than ANOVA. As the results for $\alpha = 0.1$ look very similar to the results for $\alpha = 0.05$, we did not plot them. So in general, the Tobit model is anti-conservative for our data, while the ANOVA F-test still produces reliable results. Especially for variables with a high amount of zeroes (as CHERRY1, CITRUS, ESTERY and RED.-FRUI, cf. Fig. 1), for which Guillet et al. (2001) point out the high power of the Tobit model, the empirical levels of the likelihood ratio test in the Tobit model are much too high. So the higher power of the Tobit model has a simple explanation: the associated test is highly anti-conservative.

4.3.2. Tests for equality of two products

Another question of interest is to test whether two products can be distinguished from each other. Without loss of generality, we decided to compare the products EPP2 and UCAI.

For ANOVA, we then could use the standard t-test. For the Tobit-model, we used a test statistic, where the numerator is the estimate of the difference $\tau_i - \tau_j$ and the denominator is the square root of the approximate variance of this estimate, derived from Theorem 1.

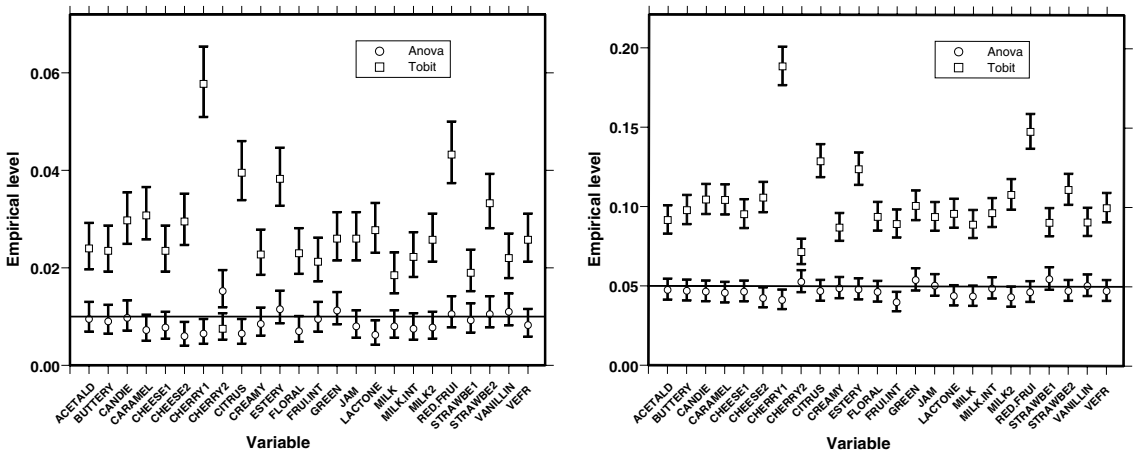

Fig. 5. Empirical levels for the F-test (ANOVA) and the χ^2-test (Tobit model) for equality of all products with 95%-confidence intervals. Left: $\alpha = 0.01$. Right: $\alpha = 0.05$.

Assuming validity of Theorem 1, this statistic is approximately standard normally distributed.

Our judgement then will be based on the same criteria as chosen in Section 4.3.1.

The results for the t-test in ANOVA are very similar to the results for the F-test. For most of the variables, the comparison graphs look like the one for the variable LACTONE (see Fig. 6), with only slight differences between the theoretical and the empirical distribution function. Most of the differences of the worst fit (for variable CHERRY2) are visible in the area of quantiles between 0.3 and 0.7 which is not of large interest for the levels usually considered. There are only slight differences in the tails. So we can expect ANOVA again to behave well for this test.

Interestingly, the empirical and theoretical distribution function for the Tobit model seem to fit better for the paired comparison than for the χ^2-test in Section 4.3.1. However, a closer look at the ACETALD attribute in the left hand side of Fig. 7 reveals that the empirical distribution function is larger than the theoretical distribution function for small values and smaller than the theoretical distribution function for large values of the test statistic. This structure can be seen for nearly all attributes—and it means that the test becomes slightly anti-conservative.

Surprisingly, the variables with a large number of zeroes, like e.g., the variable CHERRY1 in the right hand side of Fig. 7, have an entirely different structure. Here, we observe a large number of test statistics equal to zero. Closer inspection shows that this does not happen if both products have exactly the same observations, but if one of the products has only responses equal to zero.

This phenomenon can already be seen when we analyse the original un-permuted observations of CHERRY1. We find that each assessor observed less than 0.3 for the product IPIC. For the product IFMU all assessors but one gave less than 0.3, while assessor 13 gave

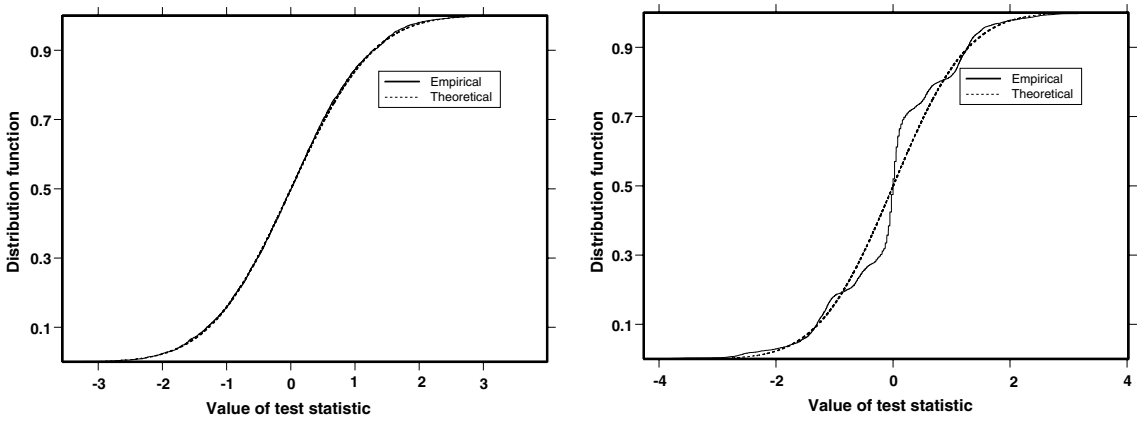

Fig. 6. Comparison of the empirical and theoretical distribution function for the test for equality of two products when using ANOVA. The theoretical distribution of the t-test statistic is the t-distribution with 264 degrees of freedom. Left: best fit for variable LACTONE. Right: worst fit for variable CHERRY2.

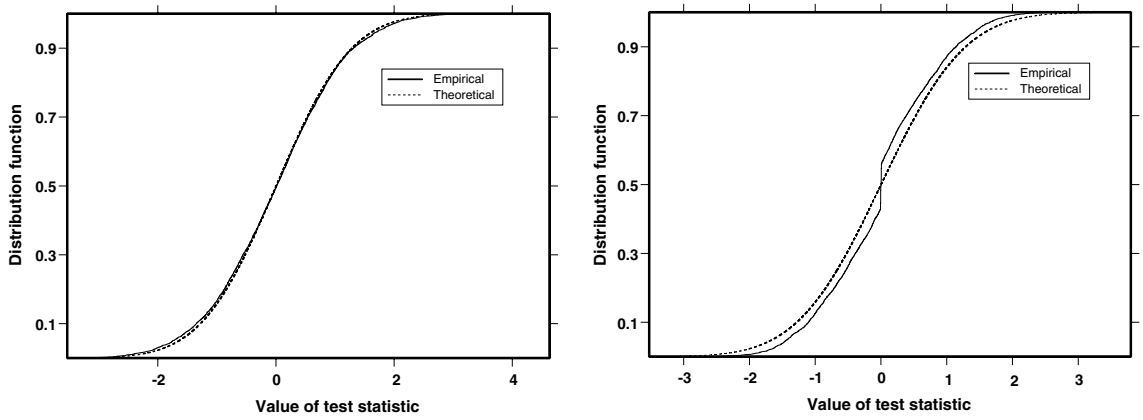

Fig. 7. Comparison of the empirical and theoretical distribution function for the test for equality of two products when using the Tobit model. The theoretical distribution of the test statistic is the normal distribution. Left: best fit for variable ACETALD. Right: worst fit for variable CHERRY1.

a 0.36 (which is just slightly above our limit 0.3). Now compare both of these two products to product AFSE which has some non-zero observations. For the comparison of IPIC and AFSE the difference of the τ_i gets estimated as -25.3 with a standard error of 25,500, resulting in a test statistic of 0.00. Comparing IFMU to AFSE, we get an estimate of -5.5 with a standard error of 2.5 resulting in a test statistic of -2.2, which is significant at the 5% level. Therefore, changing one single observation slightly (from a value just below 0.3

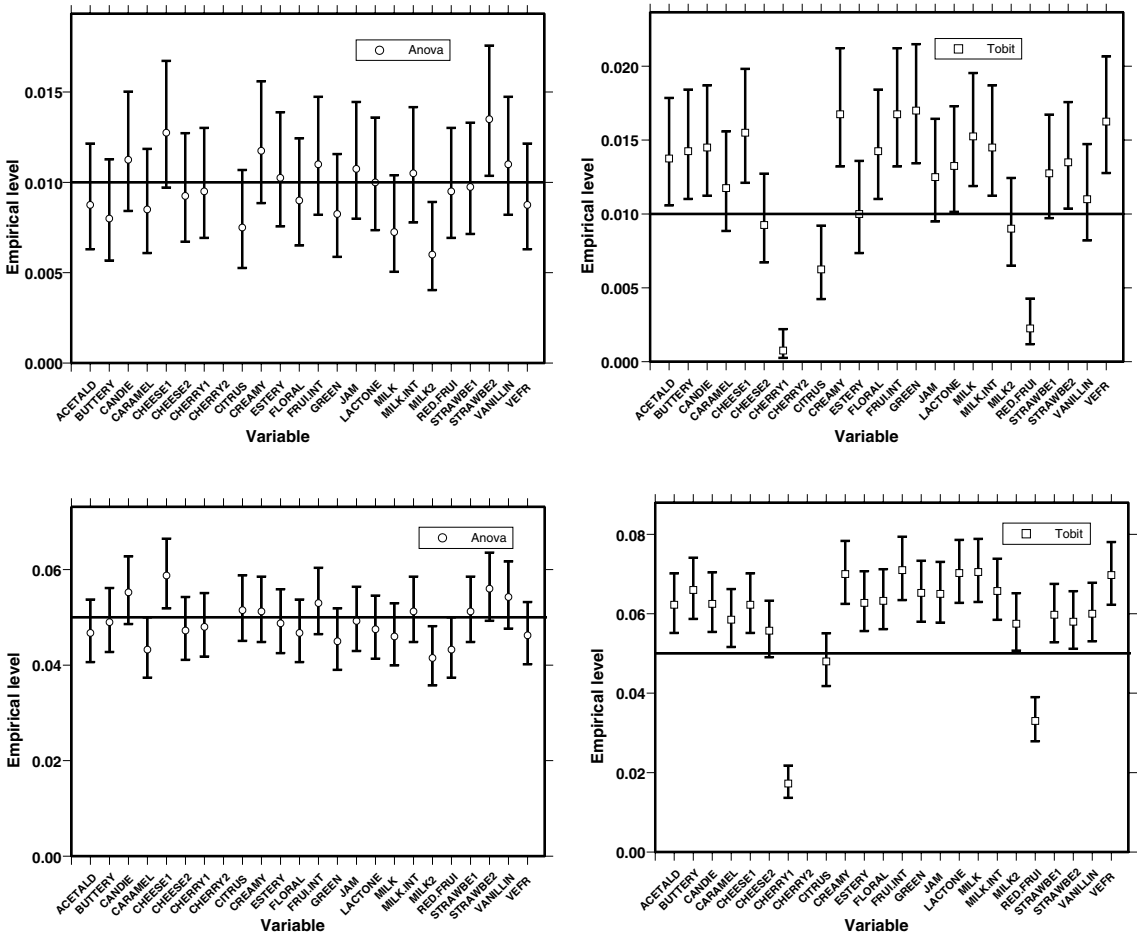

Fig. 8. Empirical levels for the tests for equality of two products with 95%-confidence intervals. Top: $\alpha = 0.01$. Bottom: $\alpha = 0.05$. Left: ANOVA t-test. Right: Tobit t-test.

to a value just above 0.3) leads to an entirely different test decision. This indicates a non-robustness of the Tobit-model that definitely is a problem.

To examine whether the tests for equality of two products keep the nominal level, we once more looked at the empirical levels with their corresponding confidence limits.

As visible in Fig. 8, the situation can be compared to the test for equality of all products. Except for very few cases, the 95% confidence intervals for the ANOVA t-tests contain the theoretical levels $\alpha = 0.01$ and $\alpha = 0.05$, respectively. Only in two cases (variable STRAWBE2 for $\alpha = 0.01$ and CHEESE1 for $\alpha = 0.05$) it might be doubted that ANOVA keeps the level. For the Tobit model, however, the majority of the confidence intervals are clearly above their respective theoretical levels. So again, the t-test for the Tobit model will lead to an anti-conservative test. For some variables with a large proportion of zeroes (CHERRY1, CITRUS and RED.FRUI), however, the test keeps the level. This is due to the phenomenon described above.

For the variable CHERRY2 with a high amount of zero values, we observed in many cases that the information matrix in Theorem 1 was not invertible. Therefore the test statistic could not be calculated, so that we omitted this variable from our considerations.

The results for $\alpha = 0.1$ are very similar to the results for $\alpha = 0.05$ and are therefore omitted.

5. Conclusions

In this article, we discuss the analysis of data from sensory profiling if there is a large number of zeroes. At the Pangborn conference in Dijon in 2001, two methods where proposed to test for differences between products for this kind of data: the well-known ANOVA model and the Tobit model which has its origin in the field of econometrics. It is possible to show for both models that their underlying assumptions are not fulfilled in the current experimental setting. For ANOVA, the normality of the data and the residuals can be questioned because of the high amount of zero data. In the Tobit model, the block structure of the Pangborn data set implies that the usual proof for the asymptotic efficiency of the estimate does not work. We performed a permutation study for the tests of equality of all products and for the tests of equality of two products. The comparison of the theoretical and empirical distribution functions and the calculation of empirical test levels both lead to the following conclusions. For ANOVA the two distribution functions agree quite well for both types of tests. There are only some slight deviations, especially for the variable CHERRY2. The Tobit model, however, shows a clearly visible distance between the empirical and theoretical distribution function for the test of equality of all products. If, however, we compare two products, then only a slight deviation can be seen. This is mainly due to the fact that in situations where one of the two products to be compared produces only observations that are zero, the test statistic becomes zero in the Tobit model.

The calculation of empirical levels strengthens these results: in general ANOVA gives tests that keep the nominal level, while the Tobit model is anti-conservative. The tests for the ANOVA model can therefore be assumed to be reliable, even though the underlying assumptions are not fulfilled. The tests for the Tobit model, however, are anti-conservative, especially when testing for equality of all products. This can be concluded from the fact that for the majority of all tests performed, the empirical level is significantly larger than the underlying theoretical level.

So we conclude that ANOVA is usable for sensory experiments, even if there is a high proportion of zeroes in the data. In contrast, the Tobit model appears too often to find differences between identical products. Therefore, we think that use of the more complicated Tobit model seems not be sensible for sensory trials, at least in its standard form. Maybe a combination of Tobit estimates with a permutation test could become useful. This, however, is a topic for further research.

References

Amemiya, T. (1973). Regression analysis when the dependent variable is truncated normal. *Econometrica, 41*, 997–1016.

Amemiya, T. (1985). *Advanced econometrics*. Oxford: Blackwell.

Bailey, R. A. (1981). A unified approach to design of experiments. *Journal of the Royal Statistical Society A, 144*, 214–223.

Conover, W. J., Johnson, M. E., & Johnson, M. M. (1981). A comparative study of tests for homogenity of variances, with applications to outer continental shelf bidding data. *Technometrics, 23*, 351–361.

Draper, N., & Smith, H. (1981). *Applied regression analysis* (2nd ed.). New York: Wiley.

Fair, R. C. (1977). A note on the computation of the Tobit estimator. *Econometrica, 45*, 1723–1727.

Guillet, M., Methot, S., & Rodrigue, N. (2001). Application of Tobit models to handle zero-valued attribute intensities. *Presented at the Pangborn conference in Dijon*.

Kunert, J., Meyners, M., & Erdbrügge, M. (2002). On the applicability of ANOVA for the analysis of sensory data. In *Proceedings 7è Journées Européennes Agro-Industrie et Méthodes Statistiques* (pp. 129–134).

Meyners, M. (2001). Anova with a large number of zeros in sensory profiling. *Presented at the Pangborn conference in Dijon*.

Olsen, R. J. (1978). Note on the uniqueness of the maximum likelihood estimator for the Tobit model. *Econometrica, 46*, 1211–1214.

R Development Core Team (2004). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. Available from: http://www.R-project.org.

Tobin, J. (1958). Estimation of relationships for limited dependent variables. *Econometrica, 26*, 24–36.