ELSEVIER

Food Quality and Preference 13 (2002) 555–559

Food Quality and Preference

www.elsevier.com/locate/foodqual

Statistical methods to examine differences in the rating of soft-drinks among different groups of consumers

Joachim Kunert*

Universität Dortmund, Fachbereich Statistik, 44221 Dortmund, Germany

Abstract

In an experiment with school children the assessors ranked five soft drinks (five brands of cola drinks) according to their preferences. An interesting aspect of the data is that the assessors could be separated into several groups: the students came from several European countries (one school each from East and from West Germany, from France, from England and from Italy). The average age of the assessors was about 15. We suppose that most statisticians would agree that comparisons among the products should be done with the Friedman test. We did not find it as obvious, however, to decide what methods we should use for the comparison among the different groups of assessors. The paper demonstrates that Hotelling's T^2-test can be used for the comparison of the groups, in spite of the fact that the observations are clearly not normal.

Keywords: Permutation test; Ranking of products; T^2-test

1. Introduction

The data for this paper were derived during a project week at a school in Dortmund (Kunert, Lehmkuhl, & Schleppe, 1995). During this week the children of this school could participate in several projects, one of which was a sensory test. We hence had a group of about 20 active pupils in our project, who helped with the technical preparation of the tests. The tests were carried out in the entrance hall of the school. The assessors were volunteers, who were asked to participate, when they passed by. They were asked to sort five soft-drinks (five brands of cola drinks) according to their preference. More precisely, each assessor received five glasses containing different brands of soft-drinks. The assessors did not know which glass contained which brand, and they were not told which brands were compared in the study. Each assessor was asked to taste from each glass at least once and then to sort the products. Hence the response from each assessor was a vector of ranks, where the least preferred drink would get a 1.

An interesting aspect of the project was the fact that there were visitors from other schools during the project week. We therefore could get the preferences of 32 school children from a school in Dortmund (which is in the western part of Germany), 19 from a school in East Germany, 21 from a school in Italy, five from a school in France and eight from a school in England. We also had a group of 24 statistics students from the department of statistics, university of Dortmund, who participated as a part of a statistics course.

There was of course no way to guarantee that the assessors were representative of their countries, or even their schools. We did, however, take care that the experiment was run in such a way that all systematic differences between the average ratings of the products were due to sensory differences between the brands of soft drinks. We also took care that the experiment was run in such a way that valid statistical tests of significance could be calculated to check whether observed differences between the groups of consumers could be explained by chance alone, or whether they proved true differences in liking between the groups.

2. Presentation in randomized order

The products had the same temperature and were presented in identical glasses. The drinks were not labelled, the identification of which product was in which glass was done with three-digit random numbers, which were different for different assessors. We did not have closed booths to prevent the assessors from

* Tel.: +49-231-755-3113; fax: +49-231-755-3454.
E-mail address: joachim.kunert@udo.edu

0950-3293/02/$ - see front matter
PII: S0950-3293(02)00039-3

communicating with each other. However, independence of the assessments of different assessors is not guaranteed by keeping them apart. If we gave the products to all assessors in the same order, then we would be liable to introduce dependence, even if we had the assessors in different continents. The order of presentation is liable to influence the rankings given by the assessor. For instance, it is very likely that the product which comes first will get a slightly better ranking than if it had come later. So, if any product always came first, then this product might get a better ranking than it deserves. We can avoid this source of bias by randomizing the order of presentation. However, if we randomize in such a way that a group of assessors must get the same order, then we introduce a dependence between their responses. Therefore, the experiment was randomized in such a way that each assessor received the products in his/her own random order. The order was derived independently for each assessor. This guarantees that if certain products get more often a high ranking than others, then this must be due to sensory differences between the products. If two products are identical in their sensory properties, then each will have the higher ranking of the two with probability 1/2, and the number of assessors who prefer product A would be a binomial with success probability 1/2. Under the global null-hypothesis that all products are identical, all possible vectors of ranks have the same probability. Note that under the global null-hypothesis the responses of the assessors are independent, even if we allowed them to communicate: if, for example, the assessors had fixed to rank the products in exactly the order in which they were presented, their responses would be independent, due to the independent randomization of the order of presentation. In practice, we ensured that the assessors did not communicate with each other during the experiment.

These properties introduced by the randomization made it possible to test the null-hypothesis that there are no sensory differences between the products. This hypothesis is usually tested with the Friedman test. Here, we clearly found that there were differences between the products. When we compared the average rankings from all assessors, we found that two brands (which are the market leaders) were significantly more liked than the other products. The clearest result, however, was that a small local brand was strongly disliked, its average rank was much smaller than that of the other brands (Kunert et al., 1995, for details).

However, we were not primarily interested in the comparison between the products. Our main objective was the comparison among the consumers from the different schools. Is the acceptance of the products generally similar in the different groups, or are there significant differences? When we first analysed the data (Kunert et al., 1995), we considered the Euclidean distances between the average rankings from the different schools and the university students. We used a permutation test to show that this difference was too large in the case of the English school to be explained by chance alone. For the other schools we did not find significant differences to the university students. We therefore concluded that there was a significant difference between the English pupils and the university students, while the differences between the university students and the other schools could be explained by chance.

We used a permutation test, because we did not want to make assumptions on the distribution of the observations. However, in reality we do have information on the distribution of the observations. For instance, we know that each response is a permutation of the numbers 1, 2, 3, 4, 5. We took pains to assure that the assessments from different assessors were independent, and we can assume that we would have obtained the same distribution of the responses if we repeated the experiment under identical conditions. We therefore can assume that the assessments within each group are identically distributed. A randomly selected pupil from a given school produces each possible ranking with a certain probability, which may depend on the school. Fisher (1971) in his criticism of nonparametric methods, claimed that it is wrong to pretend to know nothing, when we do know something. He proposed to use all information on the data that is available to us, and to do a parametric analysis.

A standard parametric method to compare two vectors of means would be to use Hotelling's T^2-test. We did not use it in Kunert et al. (1995) because we thought that the conditions for the applicability of the T^2-test would not be fulfilled.

What are the conditions of the T^2-test? The most important condition is that the vectors of responses of the assessors within each group are independent identically distributed. As pointed out before, this assumption was justified by the design of our experiment. The second condition, which is always stressed as very important, is that the observations from each assessor should be multivariate normal. This condition is clearly not fulfilled for our data, since we have discrete observations. Each vector of responses from one assessor is a permutation of the numbers 1, 2, 3, 4, 5.

3. Normality assumption

Is the normality of the single observations really necessary? It must be realized that we are comparing means of observations. Therefore, the central limit theorem is working in our favour. If we assume that the assessors in a given group are a random sample from an infinitely large set, then the p dimensional vectors $\boldsymbol{x}$ of responses from the assessors in a given group are independent identically distributed with expectation-vector $\boldsymbol{\mu}$

and covariance matrix Σ, and if this covariance matrix is invertible, then the statistic

$$\mathbf{N} = \sqrt{n}\Sigma^{-1/2}(\bar{\mathbf{x}} - \boldsymbol{\mu}),$$

for large n, is approximately multivariate normal, where the expectation of $\mathbf{N}$ is the p-vector of zeros and the covariance matrix equals the $p \times p$ identity matrix.

If we compare two groups of consumers, both of which have a large sample size, then under the hypothesis that the distributions in both samples are the same, this central limit theorem directly implies that the T^2-statistic

$$T^2 = \frac{n_1 n_2}{n_1 + n_2}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T \mathbf{S}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)\frac{n_1 + n_2 - p - 1}{p(n_1 + n_2 - 2)}$$

is approximately F-distributed with p degrees of freedom for the numerator and $n_1 + n_2 - p - 1$ degrees of freedom for the denominator. Here $\mathbf{S}$ is the empirical covariance matrix, which converges almost surely to the unknown theoretical covariance matrix Σ. Note that T^2 is only defined if $\mathbf{S}$ is invertible.

If Σ is invertible, then for $n_1 + n_2$ sufficiently large, $\mathbf{S}$ will also be invertible. Therefore, non-normality of the single observations is not a problem for the T^2-test, provided the sample sizes are sufficiently large, and the covariance matrix Σ is invertible.

This shows that the applicability of the T^2-test requires a third condition. Namely, that the covariance matrix of the observations must be of full rank. Interestingly, this condition is not fulfilled for our vectors of ranks: for each assessor the vector of ranks which he/she gives to p products adds up to the fixed number $p(p-1)/2$. This holds, whatever order the assessor might prefer. Consequently, the sums of the elements of the expectation vector $\boldsymbol{\mu}$ is $p(p-1)/2$ in any group, and the sum of the elements of the covariance matrix Σ is zero. This implies that the covariance matrix is not invertible, and the inverse of $\mathbf{S}$ does not exist.

There is a way out of this problem, which is a standard methodology, cf. Srivastava and Carter (1983, section 7.2.3). Due to the fact that the sums of the elements in the expectation vectors are fixed, there cannot be a difference between the sums of the elements of the expectation vectors from the two groups. We therefore are not interested in testing whether the sums of the elements of the expectation vectors are equal in both groups. We know that they must be equal. We would therefore only want to compare the $p-1$ dimensional vectors

$$\tilde{\boldsymbol{\mu}} = \mathbf{C}\boldsymbol{\mu} = \begin{bmatrix} \mu_1 - \mu_p \\ \mu_2 - \mu_p \\ \vdots \\ \mu_{p-1} - \mu_p \end{bmatrix},$$

and test whether there are differences in the $\tilde{\boldsymbol{\mu}}$ between the groups. This hypothesis is checked by comparing the transformed observations z, where

$$\mathbf{z} = \mathbf{C}\mathbf{x} = \begin{bmatrix} x_1 - x_p \\ x_2 - x_p \\ \vdots \\ x_{p-1} - x_p \end{bmatrix}.$$

For these vectors the covariance matrix has full rank. Therefore, the T^2-statistic can be calculated.

It is therefore possible to compare the pairs of groups of assessors with the help of the T^2-test.

4. Permutation test to demonstrate the applicability of the T^2-test

To demonstrate that the T^2-test can indeed be used for the comparison of two groups of vectors of ranks when the sample sizes are of the size considered in our experiment, I use a method which I have employed frequently, see for example, Kunert (1998, 2000), and which was proposed by Fisher in 1935 (see Fisher, 1971). We compare the empirical distribution function derived from a permutation test to the theoretical distribution function that we would get if all conditions for the application of the T^2-statistic were ideally fulfilled.

To calculate the permutation test, we consider the null-hypothesis that the distributions of the response are the same in both groups of assessors. Then the distribution of the T^2-statistic would not change if we exchanged assessors between the groups. We therefore might redistribute the assessors randomly among the two groups and calculate the fictional T^2-statistic for these rearranged assessors. We can repeat this m times and compare the truly observed T^2-statistic to the set of all fictional T^2-statistics. If the truly observed T^2-statistics is among the $\alpha \times m$ largest, then we can reject the null-hypothesis at an error level α.

As an example, we might do this for the comparison of the English school and the university students. Here we have calculated 10,000 fictional T^2-statistics. The truly observed T^2-statistic was 4.88. Note that there were $n_1 = 8$ children in the English group, while we had $n_2 = 24$ university students. The reduced vector $\mathbf{z}$ was four-dimensional.

Therefore, the T^2-test would compare the observed T^2 to the F-distribution with 4 and 27 degrees of freedom. This implies that we have a theoretical P-value which equals 0.004. In our simulations we had 40 observations out of 10,000 which gave a fictional T^2 that was larger than 4.88, hence we had an empirical P-value of exactly 0.004 (!). Fig. 1 shows that this excellent fit was received for any possible value of T^2: the

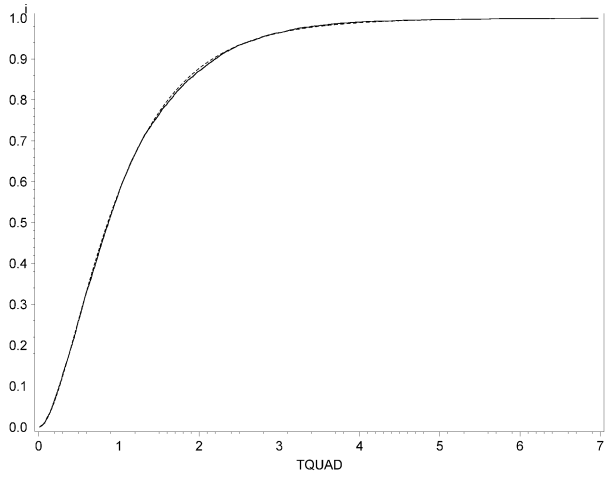

Fig. 1. Empirical distribution function of the permutation test for the comparison of the English pupils to the university students.

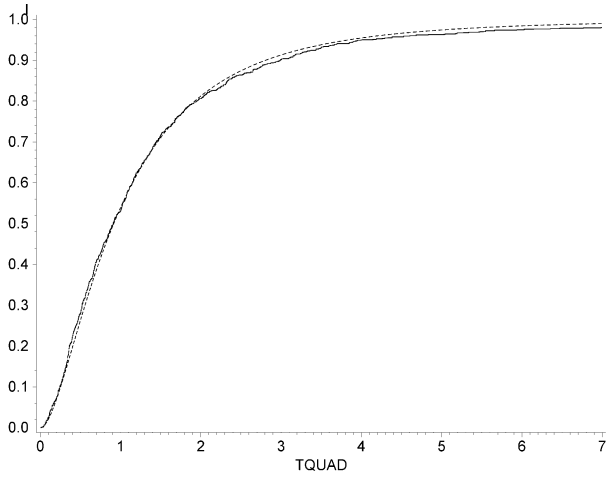

Fig. 2. Empirical distribution function of the permutation test for the comparison of the English pupils to the French pupils.

empirical distribution function of the fictional T^2 (closed line) is almost identical to the theoretical distribution function (dotted line).

The fit between the empirical and the theoretical distribution function is excellent. Note that for the applicability of the permutation test we did not have to make the assumption that each of the groups of assessors is a random sample from some infinite set. We only needed the assumption that under the null-hypothesis the distribution of the rankings is the same in both groups. The exact fit shows that the assumption that the assessors are a random sample, which was necessary for the formal proof of the central limit theorem, is not necessary for the approximate validity of the F-distribution for the T^2-statistic.

Fig. 2 shows that the fit is also sufficiently good for smaller sample sizes. If we compare the English students to the French students, then we have $n_1 = 8$ and $n_2 = 5$. Even for this relatively small size, the fit is sufficiently good.

5. Multiple testing

Note that we have a multiple test problem. If we take all pairs of schools among the six groups in the study, then we make 15 comparisons. There is a considerable risk that at least one of the pairs gives a significant difference by pure chance. We have tried to avoid this risk by two means. Firstly, we used the Bonferroni-inequality to adjust the P-values. We therefore would multiply the P-values by the number of T^2-tests calculated. Secondly, we restricted the number of T^2-tests, to avoid having to be too conservative. (The selection of the T^2-tests to do had to be done before we calculated them, of course. It would be nonsense to calculate all 15 T^2-tests, then select the five smallest P-values and multiply them by 5.)

We selected the group of the university students as a standard, which we compared to all other groups. This was logical, because we knew the number of university students (namely 24) in advance. Therefore, we could be sure that this group would be sufficiently large. Among the school children, the group from Dortmund was the largest: we had 32 West-German school children. In fact, we could have had many more, but we decided to have this group comparable in size to the others. (We therefore did a triangle test on the differences between coca cola in glass bottles and in plastic bottles with 85 other Dortmund pupils, who wanted to take part in the experiment.) The other groups were smaller. We had eight English assessors, five French, 21 Italian and 19 East German assessors. Hence the total number of assessors was 109. (We had a 110th assessor, who was a newspaper journalist, but her results were not used for the analysis.) There were about 20 pupils from England and from France visiting Dortmund during the week, but we did not reach more than eight and five, respectively. Their teachers were not as interested as the Italian and the East German teachers, who sent their pupils to participate. We therefore could only get those English and French pupils who passed by chance (like the West German pupils, but there were many more possible candidates). We kept a list of the assessors names, such that we did not have multiple assessments.

Table 1 lists the mean ranks derived by the different groups. The most striking difference is the performance of the product "Hartinger", which had the lowest mean rank in all groups, except for the English pupils, where it had the highest rank together with "River Cola". River Cola had the second lowest rank in all three German groups, while it was third for the Italian and the French group. This is interesting, because River Cola had had a very good rating in a report from the German consumer organisation (Stiftung Warentest, 1991).

Table 1
The average ranks given by the groups of assessors to soft-drinks that were presented anonymously

Group	River Cola	Classic Cola	Pepsi Cola	Hartinger	Coca Cola
English	3.4	2.6	2.4	3.4	3.2
French	3	3.2	4.2	2	2.6
West-German	2.9	3.2	3.3	1.7	3.9
Italian	3.2	3.0	3.5	1.9	3.4
University	2.5	3.7	3.5	1.9	3.5
East-German	2.6	3.3	3.7	2.1	3.4

Table 2
The T^2-tests for the comparison of the different schools with the university students

Group	T^2-statistics	P-value
English	4.88	0.0043
French	1.41	0.26
West-German	1.36	0.26
Italian	1.63	0.19
East German	0.48	0.75

Table 2 lists the T^2-statistics and the P-values for the comparison of the different schools with the university students. Note that the different T^2-statistics are compared to F-distributions with different degrees of freedom for the denominator. Hence the P-values are not necessarily strictly monotonous in the T^2-statistics.

Our finding is that there was a significant difference between the English school and the university students. The corresponding p-value is less than 0.05 even if we multiplied it by 5. All other differences could be explained by chance. The corresponding P-values are larger than 0.05, even without multiplication by 5.

We concluded that the English school was obviously different. For the other groups we think that their ratings are similar.

Acknowledgements

Financial support by the Deutsche Forschungsgemeinschaft (SFB 475) is gratefully acknowledged.

References

Fisher, R. A. (1971). *The design of experiments* (reprint of the 8th ed.). New York: Hafner Publishing Company.

Kunert, J. (1998). Sensory experiments as crossover studies. *Food Quality and Preference*, *9*, 243–253.

Kunert, J. (2000). Workshop on the statistical analysis of sensory profiling data: randomization/permutation/ANOVA. *Food Quality and Preference*, *11*, 141–143.

Kunert, J., Lehmkuhl, F., & Schleppe, A. (1995). ''Ein Statistisches Experiment mit Schülern auf Bevorzugung von Erfrischungsgetränken''. *Stochastik in der Schule*, *15*(3), 23–42

Srivastava, M. S., & Carter, E. M. (1983). *An introduction to applied multivariate statistics*. New York: North Holland.

Stiftung Warentest (German Consumer Organisation). (1991). Oft mehr Werbung als Geschmack. *Test (Journal of the German consumer organisation)*, *6*(91).