Food Quality and Preference 49 (2016) 129–139

Contents lists available at ScienceDirect

Food Quality and Preference

journal homepage: www.elsevier.com/locate/foodqual

Delta-tilde interpretation of standard linear mixed model results

Per Bruun Brockhoff [a,*], Isabel de Sousa Amorim [b], Alexandra Kuznetsova [a], Søren Bech [c], Renato Ribeiro de Lima [b]

[a] DTU Compute, Statistics Section, Technical University of Denmark, Richard Petersens Plads, Building 324, DK-2800 Kongens Lyngby, Denmark
[b] DEX – Departamento de Ciências Exatas, Universidade Federal de Lavras, Campus da UFLA – Caixa Postal 3037, Lavras, MG, Brazil
[c] Bang & Olufsen A/S, Struer and Aalborg University, Denmark

ARTICLE INFO

Article history:
Received 16 June 2015
Received in revised form 19 August 2015
Accepted 20 November 2015
Available online 22 November 2015

Keywords:
Visualization
Effect size
Analysis of variance
F test
d-Prime

ABSTRACT

We utilize the close link between Cohen's *d*, the effect size in an ANOVA framework, and the Thurstonian (Signal detection) *d*-prime to suggest better visualizations and interpretations of standard sensory and consumer data mixed model ANOVA results. The basic and straightforward idea is to interpret effects relative to the residual error and to choose the proper effect size measure. For multi-attribute bar plots of *F*-statistics this amounts, in balanced settings, to a simple transformation of the bar heights to get them transformed into depicting what can be seen as approximately the average pairwise *d*-primes between products. For extensions of such multi-attribute bar plots into more complex models, similar transformations are suggested and become more important as the transformation depends on the number of observations within factor levels, and hence makes bar heights better comparable for factors with differences in number of levels. For mixed models, where in general the relevant error terms for the fixed effects are not the pure residual error, it is suggested to base the *d*-prime-like interpretation on the residual error. The methods are illustrated on a multifactorial sensory profile data set and compared to actual *d*-prime calculations based on Thurstonian regression modeling through the ordinal package. For more challenging cases we offer a generic "plug-in" implementation of a version of the method as part of the R-package SensMixed. We discuss and clarify the bias mechanisms inherently challenging effect size measure estimates in ANOVA settings.

1. Introduction

Data analysis within the sensory and consumer science fields can be particularly challenging due to use of humans as the measurement instrument. Understanding how responses change due to product differences versus change due to subject differences is important. Analysis of variance (ANOVA) is one of the most often employed statistical tools to study differences between products when they are scored by either categorical rating (ordinal) scales and/or unstructured line scales. If for instance one finds that the main product effect is significant, one will be interested in knowing more about which products are different from each other. To complement the ANOVA *F*-table, post hoc tests are performed. These procedures, also called multiple comparison tests, are generally based on some correction to protect against having the multiple testing procedure invalidating the overall significance level. Some of the commonly used methods include the Tukey, Bonferroni Newman–Keul's and Ducan's procedures (Næs, Brockhoff, & Tomic, 2010).

Data analysis based on analysis of variance within the sensory field is usually characterized by a number of such relevant post hoc analyses. To some extend this then handles the effect interpretation part of the analysis. However, it is still valuable to be able to supplement the initial overall ANOVA *F*-testing, often with highest focus on the *p*-values with some good measures of overall effect size. In the widely used open source software PanelCheck (Mat et al., 2008) the inbuilt ANOVA results are visualized by multi-attribute bar plots of *F*-statistics combined with color coding of the significance results. In this way the *F*-statistic is used as a kind of effect size measure. This can be a good approach, especially within PanelCheck, where the multi-attribute bar plot of the overall product differences are used only for single-factor product effects and with the same choice of *F*-test denominator across all the attributes of a plot.

[*] Corresponding author.
E-mail address: perbb@dtu.dk (P.B. Brockhoff).

http://dx.doi.org/10.1016/j.foodqual.2015.11.009
0950-3293/© 2015 Elsevier Ltd. All rights reserved.

However, the F-statistic itself is generally not the best measure of effect size as it depends on the number of observations for each product. And the various ANOVA mixed models, that we often use for such analysis also complicates the relative effect size handling as generally in mixed models, different effects may have different noise structures, that is, different factors may be tested using different F-test denominators. Moreover, as was pointed out in Kuznetsova, Christensen, Bavay, and Brockhoff (2015), it is important, specifically within the sensory and consumer field to be able to also handle more complicated settings than the most simple ones.

More recently, a number of new open source software tools with, among other things, focus on more extended type of mixed model ANOVA for sensory and consumer data have appeared. The ConsumerCheck (Tomic, Brockhoff, Kuznetsova, & Næs, Submitted for publication) a tool developed in the same spirit as PanelCheck, offers quite general mixed model analysis of consumer data based on the newly developed more generic R-package lmerTest (Kuznetsova, Brockhoff, & Christensen, 2015a). In addition, in the still developing R-package SensMixed (Kuznetsova, Amorim, Christensen, Lima, & Brockhoff, 2015) one of the main purposes is to provide nice and visual multi-attribute interpretations of more complicated analyses. The resulting multi-attribute bar plots will then involve different factors with different number of levels and different number of observations within the levels. It may also involve different mixed model error terms for different factors. All of this calls for some careful thoughts on how to visualize the results of the (mixed) ANOVA results in the best possible way.

The purpose of the present study is to suggest better multi-attribute ANOVA plots for sensory and consumer data based on an effect size expressed in terms of relative pairwise comparisons. We will show how this has a close link to the Thurstonian d-prime, and as such is a generic measure that can be interpreted and compared across any attribute and situation. For balanced data settings, the measure becomes a simple transformation of either one or a few F-statistics making the approach easily applicable for anyone for these cases. For more challenging cases we offer a generic "plug-in" implementation of a version of the method as part of the R-package SensMixed (Kuznetsova, Brockhoff, & Christensen, 2015b).

The paper is organized such that first, in Section 2, we introduce the basic notion of effect size (ES) in ANOVA framework and the concepts of d-prime. Then in Section 3, we define the effect size $\tilde{\delta}$. Next, in Section 4 it is shown how to estimate the $\tilde{\delta}$ ES measure for certain relevant standard mixed models with possible bias correction. After this, in Section 5 we illustrate the method on a multifactorial sensory profile data set and compare the $\tilde{\delta}$ proposed here with the actual d-prime based on Thurstonian modeling. The paper ends with discussions in Section 6.

2. Cohen's *d* and *d*-prime – important effect size measures

Analysis of variance (ANOVA) is one of the most used and the most important methodologies when focus is on investigating product differences in sensory and consumer studies (Næs et al., 2010). ANOVA includes a particular form of null hypothesis statistical testing (NHST) used to identify and to quantify the factors that are responsible for the variability of the response. The null hypothesis for ANOVA is that the means of the factors are the same for all groups. The alternative hypothesis is that at least one mean is different from the others. An F-statistic is obtained in the ANOVA and the F distribution is used to calculate the p-value.

The NHST is a direct form and an easy way to conclude about the statistical significance of a factor, by considering a significance level and a p-value. However, it gets a lot of criticism from researchers of different fields. Yates (1951) observed that researchers paid undue attention to the results of the tests of significance and too little attention to the magnitudes of the effects, which they are estimating. NHST addresses whether observed effects stand out above sampling error by using a test statistic and its p-values, though it is not as useful for estimating the magnitude of these effects (Chow, 1996).

A similar point is made by Sun, Pan, and Wang (2010) and Cohen (1994) phrases it in the following way: "the NHST does not tell us what we want to know, and we so much want to know what we want to know, that, out of desperation, we nevertheless believe that it does!"

The ongoing debate on statistical significance tests has resulted in alternative or supplemental methods for analysing and reporting data. One of the most frequent recommendations is to consider effect size estimates to supplement p-values and to improve research interpretation (Cohen, 1990, 1992, 1994; DeVaney, 2001; Coe, 2002; Steiger, 2004; Cumming & Finch, 2005; Fan, 2010; Sun et al., 2010; Kelley & Preacher, 2012; Grissom & Kim, 2012). Cohen (1990) affirms that the purpose should be to measure the magnitude of an effect rather than simply its statistical significance; thus, reporting and interpreting the effect size is crucial. Fan (2010) shows that p-value and effect size complement each other, but they do not substitute for each other. Therefore, researchers should consider both p-value and effect size.

Cohen (1992) established a relation between the effect size (ES) and NHST definitions: the ES corresponds to the degree in which the H_0 is false, i.e., it is a measure of the discrepancy between H_0 and H_1. Grissom and Kim (2012) states that whereas a test of statistical significance is traditionally used to provide evidence (attained p-value) that the null hypothesis is wrong; an ES measures the degree to which such a null hypothesis is wrong (if it is false).

In other words, an effect size is a name given to a family of indices that measure the magnitude of a treatment effect. It can be as simple as a mean, a percentage increase, a correlation; or it may be a standardized measure of a difference, a regression weight, or the percentage of variance accounted for. For a two-group setting, the ES quantifies the size of the difference between two groups, and may therefore be said to be a true measure of the significance of the difference (Coe, 2002).

An important class of ES measures is defined by using the standardized effect size. In this class are included the Cohen's d, which is the difference measured in units of some relevant standard deviation (SD) (Cumming & Finch, 2005). Cohen's d is the ES index for the t test of the difference between independent means expressed in units of (i.e., divided by) the within-population standard deviation, which is given by

$$d = \frac{\mu_a - \mu_b}{\sigma}$$

where μ_a and μ_b are independent means and σ is the within-population standard deviation.

There are several effect size measures to use in the context of an F-test for ANOVA. Cohen (1992) defined the effect size for one-way ANOVA as the standard deviation of the K population means divided by the common within-population standard deviation:

$$f = \frac{\sigma_m}{\sigma} \tag{1}$$

where σ_m is the standard deviation of the K population means and σ is the within-population standard deviation.

A very similar measure of standard ES for ANOVA is the root-mean-square standardized effect (Ψ) presented by Steiger (2004). Considering the one-way, fixed-effects ANOVA, in which K means