Food Quality and Preference 47 (2016) 29–33

Contents lists available at ScienceDirect

Food Quality and Preference

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

Testing for differences between impact of attributes in penalty-lift analysis

Michael Meyners

Procter & Gamble Service GmbH, 65824 Schwalbach am Taunus, Germany

ARTICLE INFO

Article history:
Received 15 August 2014
Received in revised form 13 October 2014
Accepted 2 November 2014
Available online 13 November 2014

Keywords:
Penalty-lift analysis
Check-all-that-apply (CATA)
Randomization test
Parametric approximation

ABSTRACT

Penalty-lift analysis is a useful tool to identify so-called drivers of product liking or other hedonic measures from check-all-that-apply (CATA) or similar datasets that include some hedonic measures along with them. A recurrent question in recent projects is about how many attributes to consider as "top drivers", and how to define a reasonable cut-off. It is straightforward to test whether any attribute's impact is significantly different from zero, but this is likely not of major interest, as at least the top attributes will almost always show a statistically significant impact. However, how can we compare the penalty-lift of two different attributes on the hedonic response? To this end, two different strategies were considered. For the first one, the dataset is reduced to remove all information that does not really impact the difference in liking. An *F*-test from a two-way ANOVA is then employed to test whether there is a difference between the attributes' penalty-lift. As the dependency structure in a CATA study might invalidate the test, a randomization approach was used to validate the results, showing a reasonable fit of the parametric approximation. The second strategy employs a simple re-coding of the two attributes under consideration into one single variable with 4 factor levels. Using a relevant contrast from a two-way ANOVA, the impact of different attributes is compared using a *t*-test, again validated through a randomization approach. The approaches were successfully applied to amend the penalty-lift analysis for a CATA study on strawberries, in which the *t*-test proved more powerful than the *F*-test.

Introduction

Evaluation of products using a check-all-that-apply (CATA) question along with hedonic ratings on each of the products allows the data to be processed by a so-called penalty-lift analysis (PLA), as proposed for CATA data by Williams, Carr, and Popper (2011), (cf. also Meyners, Castura, & Carr, 2013; Meyners & Castura, 2014). To perform a penalty-lift analysis, the hedonic scores are averaged across all evaluations (assessors and products) in which the attribute under consideration was checked, and as well averaged across all observations for which the attribute was not checked. The difference between these two mean values provides an estimate of how much the hedonic response increases if people use the respective attribute to characterize the product they are just evaluating. If this difference is positive, a positive association between the attribute and the hedonic response is suggested, while a negative difference indicates that presence of that attribute is rather decreasing consumer *liking* of the product. Usually, the results are displayed in a bar chart, sorted from the attributes with the strongest positive impact on *liking* to the one with the strongest negative impact.

It is fairly simple to use an Analysis of Variance (ANOVA) to investigate for each of the attributes whether their impact on the hedonic response is truly different from zero. However, this is often not of primary interest, in particular not for longer CATA questions with many attributes. In these cases, we would usually expect many attributes to show a statistically significant impact. A question that arises much more frequently in our work goes as follows: Which of the attributes are the main drivers of *liking*? Looking at the bar chart depicting the penalty-lift of all attributes ordered from highest (at top) to lowest values (at bottom, cf. Fig. 1), of course the attributes on the top of this chart are the main positive drivers and those at the bottom are the main negative drivers of the hedonic response. However, how many attributes would we want to take from the top and the bottom of that chart, respectively? Some may say: "Let's just take 3 each, as that is convenient to report", but the true underlying question needs to be: For which of the topmost (bottommost) attributes is the impact seen really different, and which of these differences in impact might as well be due to chance? If *sweet* is estimated to have an impact of 3 points and *flavorsome* of 2.7 points on a 10 point scale,

E-mail address: meyners.m@pg.com

http://dx.doi.org/10.1016/j.foodqual.2014.11.001
0950-3293/© 2014 Elsevier Ltd. All rights reserved.

can we be sure that *sweet* is really the more important driver, or could it as well be *flavorsome*, and the reversal being only due to random variation?

In this paper, we will address this problem and propose two approaches to statistically test for significance of the differences in impact between attributes. The tests are based on the notion of randomization tests, but it will be shown that quick and easy parametric alternatives will provide reliable results in most cases as well.

Test proposal 1

The first attempt to address the problem uses only a subset of the data and performs a two-way ANOVA on that subset. Let PL_k denote the penalty-lift of attribute k. For any two attributes g and h, we want to test the null hypothesis $PL_g - PL_h = 0$ versus the alternative that this difference is not zero. Now there are apparently four possibilities how an assessors can elicit these two attributes for a given product:

- A: elicit neither g nor h for the product
- B: elicit g, but not h
- C: do not elicit g, but elicit h
- D: elicit both g and h for the product.

Let now $\bar{X}_i$ denote the average hedonic response across all assessors and evaluations in which they have elicited according to options A, B, C and D, respectively. Then, apparently

$$PL_g = \overline{X_B} + \overline{X_D} - \overline{X_A} - \overline{X_C} \qquad (1)$$

and

$$PL_h = \overline{X_C} + \overline{X_D} - \overline{X_A} - \overline{X_B}$$

With that, it is straightforward to find

$$PL_g - PL_h = \overline{X_B} + \overline{X_D} - \overline{X_A} - \overline{X_C} - \overline{X_C} - \overline{X_D} + \overline{X_A} + \overline{X_B} = 2 * (\overline{X_B} - \overline{X_C})$$

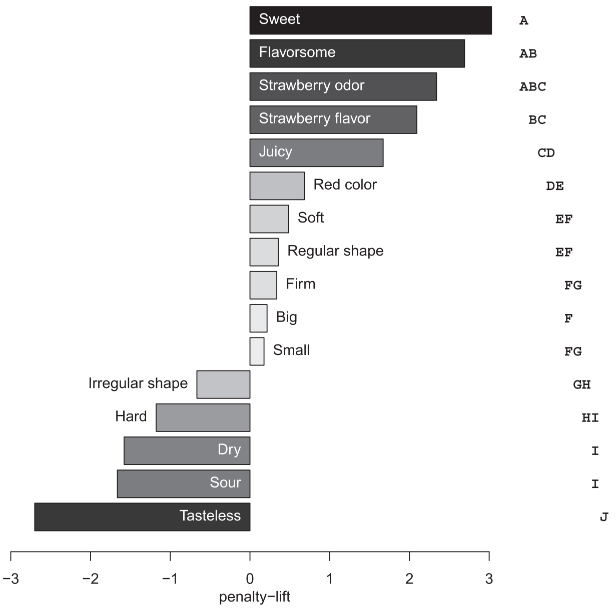

Fig. 1. Results from a penalty-lift analysis of the strawberry dataset. Penalty-lift of attributes that do not share a common letter are significantly different based on test proposal 1. The significance level used is 5%, with no correction for multiplicity being applied.

Hence, evaluations from options A and D do not contribute any information to the difference between penalty-lifts of the attributes under consideration. Therefore, these observations might as well be dropped from the dataset. Note that high counts for options A and D indicate that the two attributes are highly correlated and possibly redundant. Once done, in all remaining observations either g or h is endorsed for, but not both or none, i.e. only options B and C will still occur.

Now, some subjects may give identical information for all products, which means their data does not provide any information with regard to differences between the penalty-lift of attributes: they have endorsed attribute g for all evaluations remaining after the preceding reduction of the dataset, and never endorsed attribute h, i.e. they have opted for B throughout these observations. This in particular happens if they endorse attribute g for (almost) all products in the product category under consideration (e.g. *sweet* for lollipops), while they would never use attribute h (e.g. *spicy* in the same example). These assessors may actually bias the test results: If they give, relative to others, high *liking* scores, attribute g will get an undue high penalty-lift; it will get an undue low penalty-lift if they give below-average hedonic scores. Consequently, assessors always endorsing attribute g but never h (i.e. always using option B) contribute only to sampling variation; therefore the corresponding observations will be omitted from the evaluation. For symmetry, following the same reasoning assessors always using option C (i.e. never endorsing attribute g but always attribute h) will be omitted all the same.

It is worth noting that panelists who sometimes endorse attribute g and not h (option B), while sometimes endorsing h but not g (option C) may still bias the results to some extent if they use an extreme part of the hedonic scale: the penalty-lift as exemplified in Eq. (1) may have substantially more observations from such an assessor in option B than in C (or vice versa). Unless the hedonic scores are normalized prior to the analysis, this seems impossible to account for; in contrast to the panelists removed in the previous step (opting only for B or only for C throughout), these assessors contribute not only to the variation, but also to the information about the differences in penalty-lift of attributes, so they cannot be removed.

After this reduction of the dataset, a two-way ANOVA can be used to address the test problem, using subject and endorsement for attribute g as factors. (Of course, we could use attribute h as well, as $h = 1 - g$ holds throughout the remaining data.)

It is now important to respect the dependency structure in the data, introduced by repeated evaluations of the same assessor. Therefore, the observations are not independent. An assessor who has 5 observations in the remaining dataset, two of which endorsing for g and the other three for h, could not be considered to have endorsed four times for g and only once for h, say. Consequently, there is interdependency between the two factors in the ANOVA, which leads to the question whether p values from a parametric ANOVA (and its corresponding F-test) are valid. Therefore, the p values are determined from the classical ANOVA but as well by means of the appropriate randomization test (cf. Meyners & Pineau, 2010; Edgington & Onghena, 2007). It is worth mentioning that the number of observations that are dropped from the analysis depends on the attributes under considerations; consequently, the degrees of freedom for the respective F-distribution may vary substantially.

Test proposal 2

Is seems undesirable to omit that many observations; even though they do not appear to contribute to the difference in penalty-lift between two attributes, they might at least be used to