Bootstrapping Penalty Analysis in Sensory Evaluation of Pizza Products

Catherine Estiaga
School of Statistics
University of the Philippines Diliman

Penalty analysis is a popular method used to evaluate data from sensory evaluation using the Just About Right Scale and the Hedonic Scale. Although the test estimates the mean drops for the "Too Little" and "Too Much" categories of product attributes, penalty analysis does not provide information that can be used to test the effect of each attribute on the overall liking score. Bootstrap resampling method when used together with penalty analysis estimates the standard error of the mean drops and allows to test for the significance. This method is used in product testing of pizza products.

Key Words: bootstrap method, penalty analysis, just about right scale, hedonic scale, mean drops

1. Introduction

Penalty or Mean Drop Analysis is a method used extensively in sensory data analysis to identify potential directions for the improvement of products. It determines what attributes can lead to an increase in overall liking (Paczkowski, 2009) by showing how many points you loose for having a product "too much" or "too little" for a consumer. Penalty analysis can aid product developers in understanding differences in product preference. It can also help in understanding the basis for consumer segmentation and product substitutability (Rothman, 2007).

Product attributes used in penalty analysis are measured using the just-about-right (JAR) scales (Plaehn, 2009). JAR scale is widely used to measure the suitability of a specific attribute and to provide the optimum levels of attributes in a product. These scales are categorical variables, usually consist of five points, which assess whether there is too little, too much or just-about-right level of a product attribute (Lawless and Heyman, 1999).

Although penalty analysis has been extensively used in the food industry, there is no statistical procedure to determine the significance of the effect of each attribute on the overall liking of a product. Because the categories below and above the JAR levels are collapsed, there are not enough observations to estimate the standard error of the mean drop and to compute for the significance (Rothman, 2007). In this connection, an alternative method was developed by Xiong et al. (2007) called bootstrapping penalty analysis. Bootstrapping penalty analysis allows performing statistical testing on the results of a penalty analysis, using a technique called bootstrap to estimate variance (Xiong et al., 2007).

2. Review of Related Literature

Studies on sensory evaluation, penalty analysis, bootstrap method and bootstrapping penalty analysis are summarized below.

2.1 Sensory evaluation

Sensory Evaluation is defined by The Institute of Food Technologists (IFT) Sensory Evaluation Division USA as "a scientific discipline used to evoke, measure, analyze and interpret sensations as they are perceived by the senses of sight, smell, taste, touch and hearing" (Prell, 1976). Sensory evaluation was initially used as a service provider supplying data. However, its role has changed significantly over the years. Now, it provides insights to help guide product development, product matching and product improvement (Kemp et al., 2009). Aside from these, sensory evaluation is now widely used in process change, cost reduction, quality control, consumer acceptance, consumer preference, monitoring competition, product sensory specification, raw materials specifications, storage stability, panel selection/training and advertising claims (Stone and Sidel, 2004; Mason and Nottingham, 2002). Sensory scientists working on food industry did most of the development on sensory evaluation. It is also now being implemented by varied industries such as in personal care, paint, household cleaners, hospitality management and in many others (Chambers and Wolf, 1996).

The most widely used scale for measuring food acceptability in sensory evaluation is the hedonic scale. It was developed by David Peryam and his colleagues to assess acceptability of food items for soldiers (Peryam and Pilgrim, 1957). The hedonic scale that has been most commonly used is the 9-point scale (Table 1) in which the consumer rates their preference for food, ranging from "dislike extremely" to "like extremely" with the midpoint of 5 being "neither like nor dislike". Study showed that longer scales such as the 9-point scale tend to be more discriminating compared to a shorter one such as the 7-point scale (Jones et al., 1955).

Table 1. 9-Point Hedonic Scale

9	Like extremely
8	Like very much
7	Like moderately
6	Like slightly
5	Neither like nor dislike
4	Dislike slightly
3	Dislike moderately
2	Dislike very much
1	Dislike extremely

The JAR scale is another scale in food industries and market researchers (Lawless and Heymann, 1999) extensively used to assist them in identifying possible shortcomings of the products evaluated (Xiong et al., 2007). JAR is a dichotomous scale that measures whether a specific attribute is present in optimal levels in a product. It is usually five categories wide as shown in Table 2 and is anchored in the middle by "just right," to the top with "too much" and to the bottom with "too little" of an attribute (Stone and Sidel, 2004). Although JAR scale could have as few as three categories, this should be used carefully since rating scales usually should not have less than 5 categories (Chambers and Wolf, 1996). Aside from this, people generally tend to rate categories in the middle of the scale that is why a 5-point JAR scale is preferred over the 3-point scale (Whiston, 2009). Data gathered in this fashion are typically summarized by dividing responses to the "just right" scale into three categories: "too little (TL)," "just right (JAR)," and "too much (TM)" (Anon, 2003 and Rothman, 2007).

JAR scale provides a quick indication of attribute intensity direction. JAR data that are normally distributed around the center are indicative of an optimized level of a specific product attribute. However, information should be considered with caution when using this scale. Consumers are not generally familiar with very specific attributes so use of JAR should be limited to very simple attributes like sweetness or saltiness (Gatchalian and Brannan, 2009).

Table 2. 5-Point JAR Scale

5	Much too much
4	Too much
3	Just about right
2	Too little
1	Much too little

2.2 Penalty analysis

Penalty analysis is a test used to identify potential directions for product improvement. It is used by market researchers, sensory analysts and product developers to determine what product attributes affect most the overall liking score, purchase intent and other product-related measure (Plaehn, 2009). It assists in identifying attributes that cause an increase or decrease in hedonic scale associated with sensory attributes not at optimal levels in a product (Paczkowski, 2009) allowing the product developer to decide on what sensory properties should be improved or adjusted.

Penalty analysis itself is premised on the idea that the maximum hedonic score/overall liking will occur at the JAR point (Plaehn, 2009). Thus, it uses the data collected on the 5-point JAR scale and the liking scores on a 9-point hedonic scale.

The principle for calculating mean drops is given in Tables 3 and 4. The 5-point JAR scale for each attribute is collapsed into 3 categories – "too little," "just about right" and "too much" (Anon, 2003 and Rothman, 2007). The mean overall liking score from the 9-point hedonic scale and the percentage of respondents represented in each of the three categories are calculated. The mean drops in the overall liking score for the "too little" and "too much" from the JAR level are likewise determined (Paczkowski, 2009). Mean drops are calculated by subtracting the mean overall liking score for the JAR group to the mean overall liking score of the "too much" or "too little" categories.

Table 3. Penalty Analysis Calculation Principles (Meullenet et al., 2007)

Hedonic Scores (X)	JAR scores
7	3
6	4
7	5
8	3
9	3
6	1
5	2
2	4
4	5
6	4
6	1

Table 4. Mean Drops and % of Respondents on a Collapsed 3-point JAR Scale

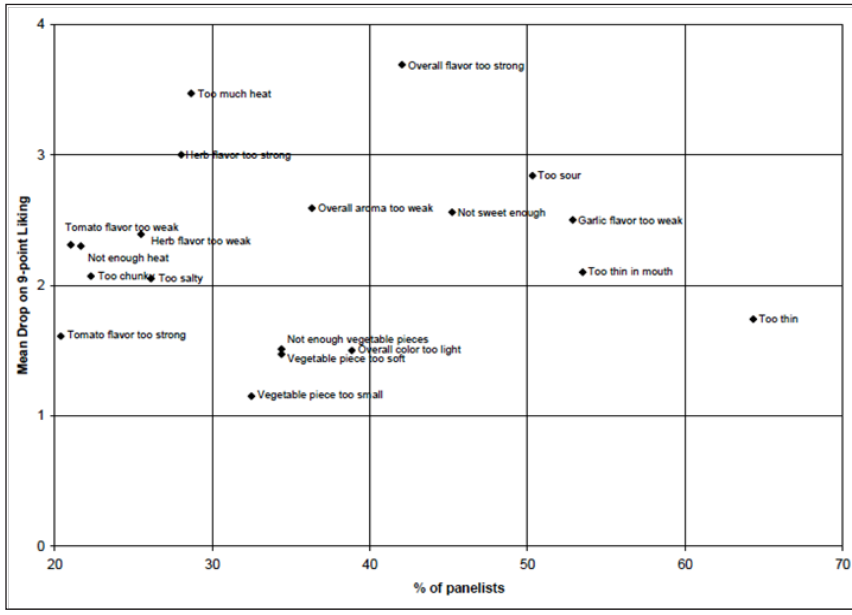
$\overline{X}_{<JAR} = 5.7$	Mean Drop $_{<JAR}$ = 8.0 – 5.7 = 2.3	% respondents $_{<JAR}$ = 3/11 = 27%
$\overline{X}_{JAR} = 8.0$		% respondents$_{JAR}$ = 28%
$\overline{X}_{>JAR} = 5.0$	Mean Drop $_{>JAR}$ = 8.0 – 5.0 = 3.0	% respondents $_{>JAR}$ = 5/11 = 45%

A minimum of 70% of the responses is usually expected in the JAR categories to conclude that a specific attribute is at its optimal level (Rothman, 2007 and Chambers et al., 1996). A skew cut-off percentage for the "too little" or "too much" categories is also advisable. This percentage depends upon the product category, sample size of the study, comfort level of the researcher and the stage of product research being conducted. Larger skews may be acceptable in earlier stages of research while a smaller skew is recommended during the later part of the study (Rothman, 2007). Since the typical skew cut-off percentage being used in the industry is 20%, mean drops are not calculated if the proportion of respondents who rated a certain attribute is less than this. Responses below 20% is too small to be considered and they might not be reliable enough (Rothman, 2007 and Meullenet et al., 2007).

Mean drops are usually represented graphically for individual products using scatter plot of the percentage of respondents in a non-JAR group on the x-axis and the mean drops on y-axis (Xiong and Meullenet, 2009) as shown in Figure 1. High negative mean drops that are associated with large proportions of respondents (upper right quadrant) are assigned greater importance than low negative or positive mean drops associated with small numbers of respondents (lower left quadrant) (Plaehn, 2009). In the figure below, the most troublesome attributes are overall flavor (too strong), garlic flavor (too weak), sweetness (not enough), overall aroma (too weak), heat (too much), herb flavor (too strong) and thickness (too thin in the mouth) (Xiong and Meullenet, 2009).

Penalty analysis helps the product developer or sensory analyst in determining what product attributes should be improved first (Meullenet et al., 2007). However, it should be noted that adjusting the JAR skews is not a guarantee that the hedonic or liking score will increase by the amount of the mean drops (Rothman, 2007).

Some advantages of penalty analysis are that it is easy to perform, is product-specific and is easily interpretable. One major limitation of this test is the fact that the categories below and above the JAR level are collapsed resulting in a loss of information. There is also no test to determine the significance of the mean drops (Xiong et al., 2007).



Source: Xiong and Meullenet (2009)

Figure 1. Example of Penalty Analysis Plot

2.3 Bootstrap method

The properties of bootstrap and its connection to the other resampling methods were not realized until Efron and Tibshirani (1993). Since then, the bootstrap has provided a powerful set of solutions for statisticians and a source of theoretical and methodological problems for statistics (Davison et al., 1986).

The bootstrap method is a well-established computer-intensive Monte Carlo technique (Xiong and Meullenet, 2009). It replicates the original data to simulate a larger population, thus allowing many samples to be drawn and statistical tests such as bootstrap variances, distributions and confidence intervals to be calculated (Chernick, 1999 and Stine, 1989).

Bootstrap can be useful in areas where it is difficult to obtain large samples (Fan, 1994). This method also does not require the assumption that the standard errors be randomly and normally distributed. Bootstrap method likewise provides a user-friendly alternative to cross-validation and jackknife to augment statistical significance testing.

The bootstrap approach consists of drawing many independent random samples by simple random sampling with replacement, evaluating the sample statistics of the corresponding bootstrap replications, and estimating the standard error of the empirical probability distribution of the data ($\hat{F}$) by the empirical standard deviations of the replications (Efron and Tibshirani, 1993; Chernick, 1999; and Diaconis and Efron, 1983).

Given $x = (x_1, x_2, \ldots, x_n)$, a random sample from x with unknown distribution F (Chernick, 1999). Let $\hat{F}$ be the empirical probability distribution of the data. An equal probability of 1/n is given on each x_i and let $x_1^*, x_2^*, \ldots, x_n^*$ be a random sample from $\hat{F}$ (Efron and Gong, 1983),

$$x_1^*, x_2^*, \ldots x_n^* \sim \hat{F} \tag{1}$$

x_i^* is not the actual data set, but rather a randomized or resampled version of x (Efron and Tibshirani, 1993). Each x_i^* is drawn independently with replacement from the original set $\{x_1, x_2, \ldots, x_n\}$ (Efron and Gong, 1983). The number of bootstrap replications required to create an ideal bootstrap sampling distribution ($\hat{F}$) is n^n where n is the original sample size. However, Diaconis and Efron (1983) and Stine (1989) have demonstrated that an ideal $\hat{F}$ can be as low as 100 replications. But for more complicated procedures for construction of confidence intervals, it has been suggested that as many as 1000 replications may be needed.

2.4 Bootstrapping penalty analysis

Bootstrapping penalty analysis allows performing statistical testing on the results of a penalty analysis using the bootstrap method to estimate variance (Xiong et al., 2007). The original data set contains the overall liking score and the JAR scale for each attribute. Mean drops for the TL and TM levels are calculated for each attribute according to the penalty analysis methodology. The bootstrap estimate of variability is obtained through resampling of the data pairs an enormous number of times (Xiong and Meullenet, 2007). The bootstrap method resamples the original data with replacement (Plaehn and Horne, 2008). A data pair is sampled and then returned to the dataset and has 1/n chance of being drawn again. All bootstrap samples have the same size as the original data. A particular pair can also appear in several of the bootstrap samples and may appear more than once in a particular sample (Xiong et al., 2007). The process of estimating the standard error of a mean drop (Meullenet et al., 2007) for a single attribute is illustrated in Figure 2.

Mean drops for the TL and TM levels are calculated for all the bootstrap samples for each attribute and are averaged to obtain the final mean drops for the two categories. The bootstrap mean and standard error of the mean are then computed based on the bootstrap replications of the mean drops (Xiong et al., 2007). Afterwards, the bias in the bootstrap estimate of the mean is removed to obtain the adjusted bootstrap mean. T-test is finally computed to determine significance of the mean drops for each attribute (Figure 2).

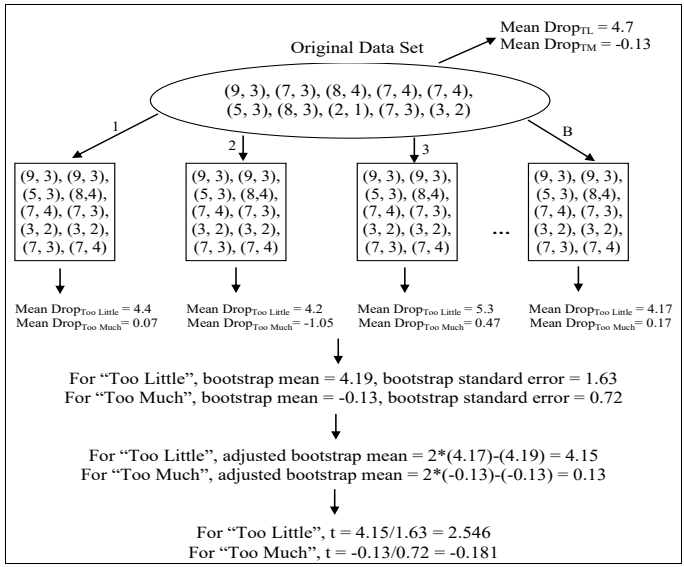


Figure 2. Scheme for Bootstrapping Penalty Analysis

3. Methodology

Sensory evaluation of two Hawaiian Pizza was done. Data gathered from the test was analyzed using bootstrapping penalty analysis. Detailed methodology is described below.

3.1 Sensory evaluation

The data used in this study was obtained from the results of in-house sensory evaluation of two Hawaiian Pizza. The two pizza products were subjected to sensory evaluation protocol by Gatchalian (1989) using 30 respondents for each pizza. Based on the studies conducted by Gatchalian and Brannan (2009) and Chambers and Wolf (1996), the number of respondents for in-house sensory evaluation can be as few as 16-20. However, the usual practice is to require at least 30 respondents. A sample size of at least 30 is considered adequate although larger sample size is much better (Levine and Stephan, 2010). The sensory evaluation conducted was only at the product development stage which maybe different when you do similar test in a pre-commercialization stage.

Respondents were pre-recruited from different areas in Metro Manila and were invited to the Product Development Center. Recruitment of respondents was carried out in advance to ensure that no seats were left unfilled and that respondents arrived on time. Respondents were from the Broad C (C1/C2) households with ages ranging from 18-35 years old. Hawaiian Pizza A has 47% female and 53% male respondents. Hawaiian Pizza B, on the other hand, has 50% female and 50%

male respondents. All respondents should have consumed Hawaiian Pizza for at least once in the past four weeks.

Hawaiian Pizza A was made of pizza sauce, cheese, 1.33 in^2 spiced ham and glazed pineapple tidbits. Hawaiian Pizza B, on the other hand, was made of pizza sauce, cheese, $1in^2$ spiced ham and fresh pineapple tidbits. The two pizzas were prepared using a family (12 inches) thick crust. After cooking, the baked pizza was sliced into eight pieces. Each respondent was given a slice of pizza for evaluation.

A proto-monadic design was used for each pizza. In the proto-monadic design, each respondent evaluates both products. Specifically, each respondent rates the first product with a monadic test followed by a preference test. The order of the products evaluated is rotated to reduce bias (Baosheng, 2005). In this study, Hawaiian Pizza A was evaluated first with a monadic test. The products were assessed in terms of overall liking using the 9-point hedonic scale, with 9 = like extremely, 1 = dislike extremely (Amerine et al., 1965). A JAR scale was also used to evaluate size of the toppings, amount of cheese, amount of meat, amount of pineapple, saltiness, cheesiness, meaty taste, pineapple taste and overall flavor blend (Table 5). The JAR scale is a 5-pointscale with the center point being just about right, the left anchor being too little of the attribute, and the right anchor being too much of the attribute (Meilgaard et al., 1999). After the monadic test, Hawaiian Pizza A and Hawaiian B were evaluated sequentially by the respondents. They were then asked which of the two pizzas they prefer more. During the test, respondents were provided with mineral water and unsalted crackers for palate cleansing. They were provided afterwards with gift certificates for their participation. The same procedure was done for Hawaiian Pizza B using 30 different respondents.

Table 5. Scales Used in the Sensory Evaluation of Hawaiian Pizza

Overall Acceptability Like extremely Like very much Like moderately Like slightly Neither like nor dislike Dislike slightly Dislike moderately Dislike very much Dislike extremely	**Saltiness** Definitely too salty Somewhat salty Just right Somewhat lacking in saltiness Not salty at all
Size of the Toppings Definitely too big Somewhat big Just the right size Somewhat small Definitely too small	**Cheesiness** Definitely too cheesy Somewhat cheesy Just right Somewhat lacking in cheesiness Not cheesy at all
Amount of Cheese Definitely too much Somewhat too much Just the right amount Somewhat too little Definitely too much	**Meaty Taste** Definitely too strong Somewhat strong Just right Somewhat weak Definitely too weak

Amount of Meat Definitely too much Somewhat too much Just the right amount Somewhat too little Definitely too much	**Pineapple Taste** Definitely too strong Somewhat strong Just right Somewhat weak Definitely too weak
Amount of Pineapple Definitely too much Somewhat too much Just the right amount Somewhat too little Definitely too much	**Overall Flavor Blend** Definitely too strong Somewhat strong Just right Somewhat weak Definitely too weak

In analyzing the results of sensory evaluation, the percentage of the top three categories in the 9-point hedonic scale was obtained for each pizza. At least 80% is needed to conclude that a product is acceptable. The result of the preference test was also counted. The minimum number of agreeing judgments is 21 to conclude that a product is significantly preferred over the other at 5% level of significance (Stone and Sidel, 2004).

3.2 Bootstrapping penalty analysis

The responses of the respondents in the 5-point JAR scale were first grouped into either too much, JAR or too little categories. The bottom two-box was categorized as "too little" of the attribute, the middle box as "JAR" and the top two-box as "too much" of the attribute. Collapsing the categories below and above the JAR level is necessary because the number of responses in the TL or TM category is often not large enough (Meullenet et al., 2007). The frequencies and percentage of respondents corresponding to the mean liking scores for the three categories were then calculated for each attribute. Mean drops for the TM and TL categories were afterwards determined by using the formula below (Xiong et al., 2007):

$$\text{Mean Drops}_{\text{Too Little}} = \overline{X}_{JAR} - \overline{X}_{\text{Too Little}} \quad (2)$$

$$\text{Mean Drops}_{\text{Too Much}} = \overline{X}_{JAR} - \overline{X}_{\text{Too Much}} \quad (3)$$

where X = overall liking score and $\overline{X}$ = mean overall liking score.

Mean drops are not computed if the percentage of respondents is less than 20% because it is too small to be considered and they might not be reliable enough. 20% is the skew cut-off percentage used in this study since it is the most commonly used in the industry (Rothman, 2007 and Meullenet et al., 2007).

A random number generator to create bootstrap samples was then done for each attribute. For each replication, a sample was selected from the original data set randomly. A data pair was obtained and then returned to the dataset and has 1/n chance of being drawn again (Xiong and Meullenet, 2007).

100 bootstrap samples were simulated from the original data. Based on the study conducted by Tibshirani (1985) and Schmidheiny (2010), B = 100 is usually adequate for estimating variances, biases and standard errors (Stine, 1989). For each attribute, mean drops and percentage of respondents for the TL and TM categories were then calculated for each of the 100 bootstrap replicates (Xiong et al., 2007) using the formulas given below.

$$S^*_{iTL} = \overline{X}^*_{iJAR} - \overline{X}^*_{iTL} \tag{4}$$

$$S^*_{iTM} = \overline{X}^*_{iJAR} - \overline{X}^*_{iTM} \tag{5}$$

where

i = 1, ..., B where B is the number of bootstrap replicates

x_i^* = $(x_1^*, x_2^*, \ldots, x_B^*)$ is a bootstrap sample drawn from the original data set

$\overline{X}^*_{iJAR}$ = mean overall liking calculated from X_i^* for the JAR Level

$\overline{X}^*_{iTM}$ =mean overall liking calculated from X_i^* for the TM Level

$\overline{X}^*_{iTL}$ = mean overall liking calculated from X_i^* for the TL Level

$S^*_{i\,TL}$ = mean drops calculated from X_i^* for the Too Little Level

$S^*_{i\,TM}$ = mean drops calculated from X_i^* for the Too Much Level

Mean drops for each of the bootstrap replicates were afterwards averaged to obtain the mean drops for the TL and TM levels for each attribute. The mean and standard error of the mean drops were computed using the formulas below (Xiong et al., 2007). By generating a large number of bootstrap samples, each observation will likely to contribute to the final variance (Plaehn and Horne, 2008).

$$\overline{s}^*_{bTL} = \frac{\sum_{i=1}^{B} s^*_{iTL}}{B} \tag{6}$$

$$\overline{s}^*_{bTM} = \frac{\sum_{i=1}^{B} s^*_{iTM}}{B} \tag{7}$$

$$\hat{s}e_{bTL} = \sqrt{\frac{\sum_{i=1}^{B}\left(s^*_{iTL} - \overline{s}^*_{bTL}\right)^2}{B-1}} \tag{8}$$

$$\hat{s}e_{bTM} = \sqrt{\frac{\sum_{i=1}^{B}\left(s^{*}_{iTM} - \overline{s}^{*}_{bTM}\right)^{2}}{B-1}} \tag{9}$$

where

B = number of bootstrap samples

$\overline{s}^{*}_{bTL}$ = bootstrap estimates of the true mean drops for the TL level

$\overline{s}^{*}_{bTM}$ = bootstrap estimates of the true mean drops for the TM level

$\hat{s}e_{bTL}$ = bootstrap estimates of the standard errors of the mean drops for the TL level

$\hat{s}e_{bTM}$ = bootstrap estimates of the standard errors of the mean drops for the TM level

To obtain the adjusted bootstrap mean, the bias in the bootstrap estimate of the mean was removed using the following (Meullenet et al., 2007):

$$\overline{s}_{bTL} = 2S_{nTL} - \overline{s}^{*}_{bTL} \tag{10}$$

$$\overline{s}_{bTM} = 2S_{nTM} - \overline{s}^{*}_{bTM} \tag{11}$$

where

S_{nTL} = original mean drops obtained from the original data set for the TL level

S_{nTM} = original mean drops obtained from the original data set for the TM level

$\overline{s}_{bTL}$ = adjusted bootstrap estimates of the means for the TL level

$\overline{s}_{bTM}$ = adjusted bootstrap estimates of the means for the TM level

After computing the standard error and the adjusted bootstrap mean, t-test was done to determine the significance of the mean drops for each attribute (Horne and Plaehn, 2008) and to determine which attributes significantly affect the overall liking of the product.

Mean drop testing was done if the number of responses on a certain attribute is greater than 20%. Penalty test was performed to test the following hypothesis:

Ho: $\mu_{JAR} - \mu_{TM} = 0$ Ho: $\mu_{JAR} - \mu_{TL} = 0$

Ha: $\mu_{JAR} - \mu_{TM} \neq 0$ Ha: $\mu_{JAR} - \mu_{TL} \neq 0$

where μ is the population mean of the TL, JAR and TM categories.

If Ho is not rejected, the mean drops on overall liking are not significant for a specific product attribute at 5% level of significance. This means that the current formulation is ready for commercialization and launching. On the other hand, if

Ho is rejected, it means that the mean drops are significant and the formulation still needs to be adjusted to obtain acceptable product.

Results of bootstrapping penalty analysis were represented graphically by plotting the mean drops on the y-axis against the percentage of respondents on the x-axis. Attributes on the upper right quadrant are the ones that need to be prioritized for improvement (Xiong et al., 2007).The formulation that will be commercialized and launched will be based not only on the results of bootstrapping penalty analysis but also on the results of overall acceptability percentage and preference test.

4. Results and Discussion

Sensory evaluation of two Hawaiian Pizza was done. Hawaiian Pizza A was prepared using pizza sauce, cheese, 1.33 in^2 spiced ham and glazed pineapple tidbits. Hawaiian Pizza B, on the other hand, was prepared using combination of pizza sauce, cheese, 1.0 in^2 spiced ham and fresh pineapple tidbits. Data from the sensory evaluation were analyzed using ordinary penalty analysis and bootstrapping penalty analysis. Results were discussed below.

4.1 Sensory evaluation

The results of sensory evaluation of Hawaiian Pizza A and Hawaiian Pizza B are shown below in Figure 3. Hawaiian Pizza B obtained a higher overall acceptability rating compared to Hawaiian Pizza A. In terms of overall acceptability, Hawaiian Pizza A and Hawaiian B were rated by most of the panelists as "like moderately" and "like very much" respectively. Overall flavor blend, on the other hand, was perceived as "somewhat too strong" for both pizza products. All other attributes were rated as "just about right" for both products.

The percentages on the top three categories of the 9-point hedonic scale was computed to obtain the overall acceptability percentage. Hawaiian Pizza A and Hawaiian Pizza B obtained an overall acceptability percentage of 86.21% and 90.32% respectively. Both are above the passing rate which is 80%. Results of preference test also showed that there is no significant preference between Hawaiian Pizza A and Hawaiian Pizza B at 5% level of significance.

Table 6. Overall Acceptability Percentage and Preference Test of Hawaiian Pizza A and Hawaiian Pizza B

	Overall Acceptability Percentage (Top three categories)	Preference Test
Hawaiian Pizza A	86.21%	13
Hawaiian Pizza B	90.32%	17

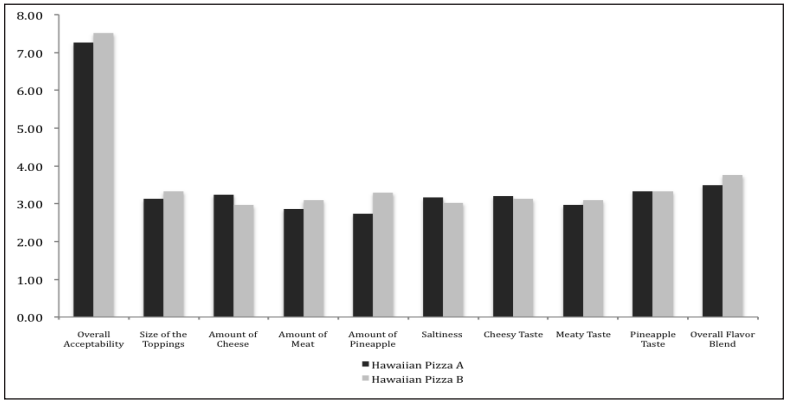


Figure 3. Results of Sensory Evaluation of Hawaiian Pizza A and Hawaiian Pizza B

4.2 Ordinary penalty analysis

Results of sensory evaluation of Hawaiian Pizza A and Hawaiian Pizza B were analyzed using ordinary penalty analysis. Figures 4 and 5 show the percentage of respondents for the too little, just about right and too much levels for the size of the toppings, amount of cheese, amount of meat, amount of pineapple, saltiness, cheesiness, meaty taste, pineapple taste and overall flavor blend of the Hawaiian Pizza A and Hawaiian Pizza B respectively.

Most of the respondents rated Hawaiian Pizza A (Figure 4) as being in the just about right level for the size of the toppings (79%), amount of cheese (54%), amount of meat (46%), amount of pineapple (64%), saltiness (82%), cheesiness (54%), meaty taste (57%),pineapple taste (61%) and overall flavor blend (43%). However, at least 70% of the responses should be at the JAR category to conclude that a specific attribute is at its optimal level. (Xiong et al.. 2007). For Hawaiian Pizza A, only two attributes namely size of the toppings and saltiness are at its optimal level.

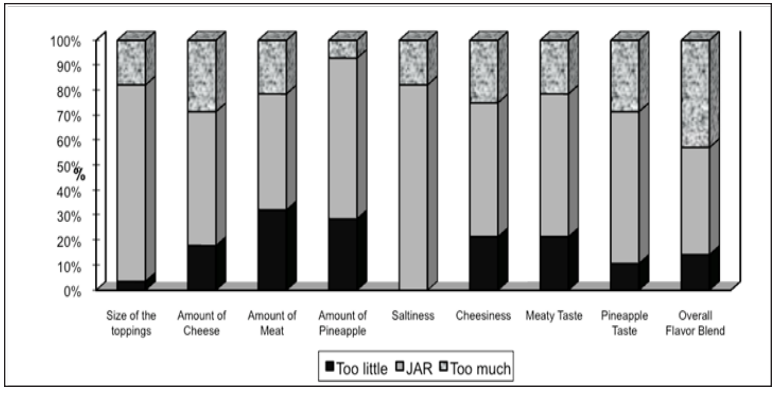


Figure 4. Percentage of Respondents for the Too Little, Just About Right and Too Much Levels of Hawaiian Pizza A

Hawaiian Pizza B (Figure 5) was also rated by most of the respondents as being in the JAR level for the size of the toppings (53%), amount of cheese (47%), amount of meat (50%), amount of pineapple (63%), saltiness (73%), cheesiness (63%), meaty taste (73%) and pineapple taste (53%). Overall flavor blend, on the other hand, was perceived by 60% of the respondents as being in the "too much" level. Saltiness and meaty taste are the only attributes of Hawaiian Pizza B that are at its optimal level.

Figure 5. Percentage of Respondents for the Too Little, Just About Right and Too Much Levels of Hawaiian Pizza B

Figures 6 and 7 show the penalty analysis plot for Hawaiian Pizza A and Hawaiian Pizza B respectively. Mean drops are represented by plotting the mean drops on the y-axis against the percentage of respondents in a non-JAR group on the x-axis. Attributes located in the upper right quadrant would be selected as those needing to be improved (Meullenet et al., 2007). Based on Figure 6, the most "troublesome" attribute is the "too little" amount of meat. 32% of the respondents considered Hawaiian Pizza A to be having too little amount of meat with a mean drop of 1.33 from the overall liking score. 29% of the respondents also considered the product to have "too little" amount of pineapple with a mean drop of 1.26. Other attributes that caused negative mean drops from the overall liking score are the "too weak" cheesiness, "too weak" meaty taste, "too much" amount of meat and "too strong" pineapple taste. Overall flavor blend and meaty taste, on the other hand, obtained positive mean drops indicating that the respondents who rated Hawaiian Pizza A as having too strong overall flavor blend and too strong meaty taste liked it more than the respondents who rated the product as "just about right."

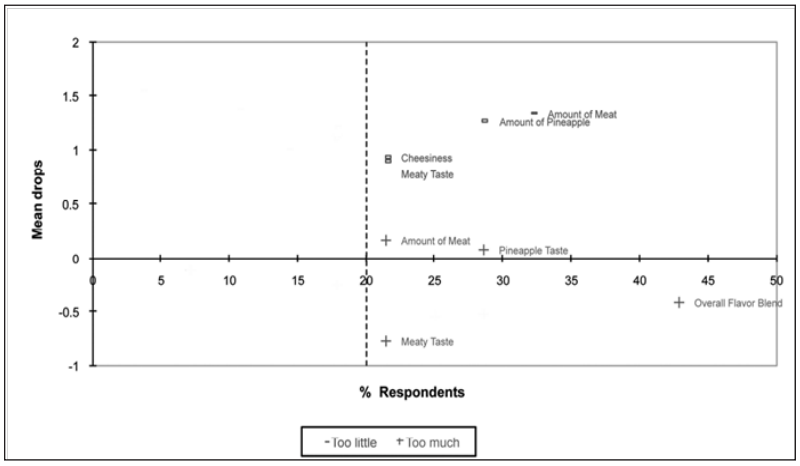


Figure 6. Penalty Analysis Plot for Hawaiian Pizza A

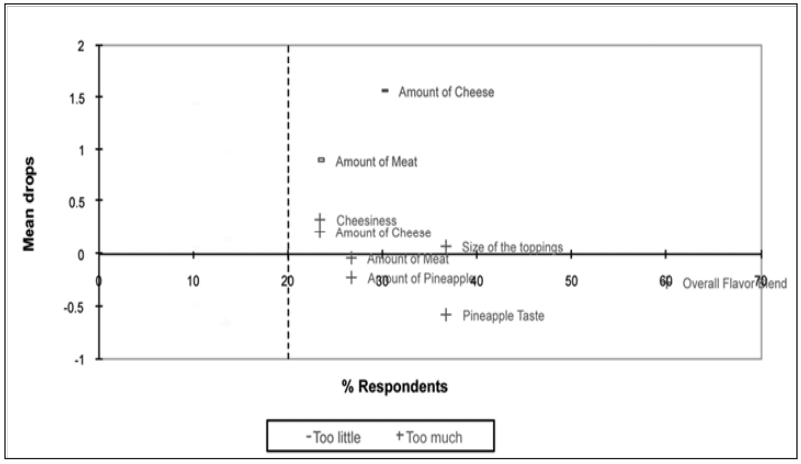


Figure 7. Penalty Analysis Plot for Hawaiian Pizza B

Figure 7 shows that the amount of meat, amount of pineapple, overall flavor blend and pineapple taste were rated as too strong by more than 20% of the respondents causing a positive mean drop on the overall liking score of Hawaiian Pizza B. These respondents liked more the said attributes than the respondents who rated those at the JAR level. Based on Figure 7, the most troublesome attribute is the amount of cheese. It was found to be "too little" by 30% of the respondents and was responsible for a large drop of 1.56 points in the overall liking score. Other attributes that are responsible for negative mean drops are too little amount of meat, too strong cheesiness, too much amount of cheese and too big size of the toppings.

4.3 Bootstrapping penalty analysis

Results of sensory evaluation of Hawaiian Pizza A and Hawaiian Pizza B were subjected to bootstrapping penalty analysis using 100 samples. Figure 8 and 9 show the percentage of respondents for the too little, JAR and too much levels for the size of the toppings, amount of cheese, amount of meat, amount of pineapple, saltiness, cheesiness, meaty taste, pineapple taste and overall flavor blend of the Hawaiian Pizza A and Hawaiian Pizza B respectively. For Figure 8, it can be observed that most of the respondents perceived the product to be in the JAR level for all attributes. However, only saltiness and size of the toppings are at its optimal level with 73% of the respondents rating the product to be in the JAR level. Most of the respondents also perceived Hawaiian Pizza B to be in the JAR level for all attributes except for overall flavor blend which was perceived as being in the "too much" level by 61% of the respondents (Figure 9). Attributes of Hawaiian Pizza B that are at its optimal level are the meaty taste and saltiness.

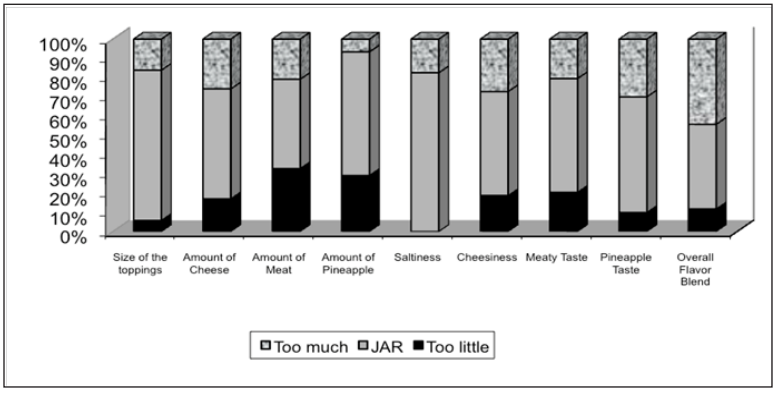


Figure 8. Percentages of Respondents for the Too Little, JAR and Too Much Levels of Hawaiian Pizza A

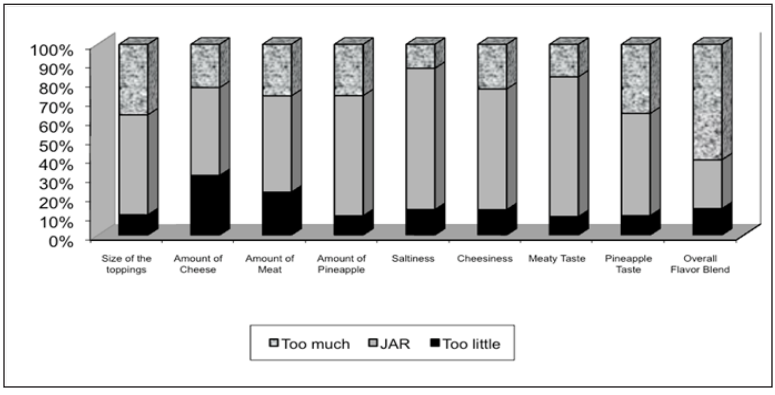


Figure 9. Percentages of Respondents for the Too Little, JAR and Too Much Levels of Hawaiian Pizza B

Figure 10 shows the penalty analysis plot for Hawaiian Pizza A. It can be observed that overall flavor blend, cheesiness, amount of cheese and meaty taste obtained positive mean drops in the overall liking score. This means that more than 20% of the respondents liked more the product attributes stated above for being in the too strong level compared to the respondents who rated those attributes in the JAR level.

Based on the Figure 11, the most "troublesome" attribute is the "too little" amount of meat. 33% of the respondents considered Hawaiian Pizza A to be having too little amount of meat with a mean drop of 1.23 from the overall liking score. Other attributes that are responsible for negative mean drops are the too little amount of pineapple, too weak meaty taste, too much amount of meat and too strong pineapple taste.

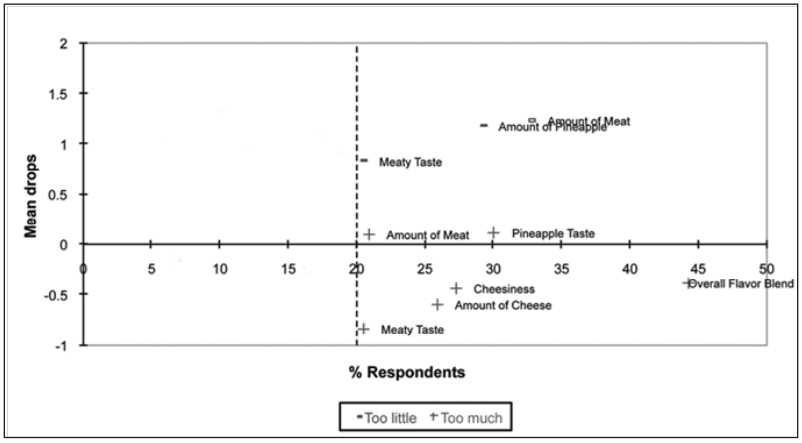


Figure 10. Penalty Analysis Plot for Hawaiian Pizza A

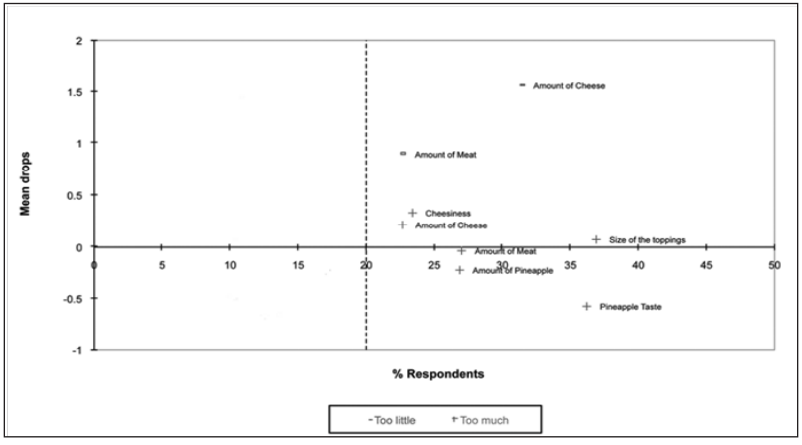


Figure 11. Penalty Analysis Plot for Hawaiian Pizza B

Figure 11 shows the penalty analysis plot for Hawaiian Pizza B. More than 20% of the respondents rated the product as having too strong pineapple taste, too much amount of meat and too much amount of pineapple. These attributes are the ones responsible for the positive mean drops on the overall liking score. The "too little" amount of cheese, on the other hand, was found to be the most troublesome attribute by 31% of the respondents and was responsible for a mean drop of 1.56. Other attributes causing negative mean drops are the too little amount of meat, too strong cheesiness, too much amount of cheese and too big size of the toppings.

T-test was done to determine the significance of the mean drops for each attribute. Tables 7 and 8 summarized the results of bootstrapping penalty analysis for Hawaiian Pizza A and Hawaiian Pizza B respectively. The results include the % of respondents for the non-JAR level, bootstrap mean drop, adjusted bootstrap mean drop, bootstrap standard error and calculated *t*.

Table 7. Results of Bootstrapping Penalty Analysis for Hawaiian Pizza A

Attributes	% Respondents	Mean Drop (Bootstrap)	Adjusted Mean Drop (Bootstrap)	Standard Error (Bootstrap)	t
Size of the Toppings					
Definitely too small	5.77				
Definitely too big	16.13				
Amount of Cheese					
Definitely too little					
Definitely too much	25.90	0.06	-1.25	0.41	1.31
Amount of Meat					
Definitely too little	32.67	0.01	2.25	0.41	-2.44*
Definitely too much	20.90	-0.01	0.20	0.50	-0.20
Amount of Pineapple					
Definitely too little	29.10	-0.03	2.38	0.35	-2.41*
Definitely too much	6.67				
Saltiness					
Not salty at all	0				
Definitely too salty	17.50				
Cheesiness					
Definitely too weak	18.67				
Definitely too strong	27.27	-0.02	-0.86	0.41	0.84
Meaty Taste					
Definitely too weak	20.37	0.01	1.66	0.51	-1.65
Definitely too strong	20.50	-0.02	-1.65	0.53	1.63
Pineapple Taste					
Definitely too weak	9.87				
Definitely too strong	30.00	0.07	0.15	0.39	0.08
Overall Flavor Blend					
Definitely too weak	11.70				
Definitely too strong	44.27	0.01	-0.78	0.37	0.79

**significant at α = 0.05*

Based on Table 7, the respondents strongly penalized Hawaiian Pizza A when they perceived the product as having definitely too little amount of meat and definitely too little amount of pineapple. The tests are significantly different at 5% level of significance and the overall liking score is reduced by 2.25 for the amount of meat and 2.38 for the amount of pineapple. Increasing the amount of meat and the amount of pineapple may cause an increase in overall liking of the product. The penalty test is not significant when the respondents perceived Hawaiian Pizza A as having definitely too much amount of cheese, definitely too much meat, definitely too much pineapple, definitely too strong cheesiness, definitely too strong pineapple taste and definitely too strong overall flavor blend. For all the other attributes, the penalty test was not computed since the percentages of respondents who rated the product in the non-JAR level is less than 20% (Xiong et al., 2007).

Table 8. Results of Bootstrapping Penalty Analysis for Hawaiian Pizza B

Attributes	% Respondents	Mean Drop (Bootstrap)	Adjusted Mean Drop (Bootstrap)	Standard Error (Bootstrap)	t
Size of the Toppings					
Definitely too small	10.73				
Definitely too big	36.90	-0.07	0.22	0.42	0.64
Amount of Cheese					
Definitely too little	31.33	-0.02	3.14	0.44	-2.21*
Definitely too much	22.67	-0.06	0.49	0.53	1.31
Amount of Meat					
Definitely too little	22.60	-0.08	1.87	0.55	-2.44*
Definitely too much	27.00	-0.04	-0.03	0.49	-0.20
Amount of Pineapple					
Definitely too little	10.23				
Definitely too much	26.87	-0.01	-0.42	0.45	0.14
Saltiness					
Not salty at all	13.47				
Definitely too salty	12.57				
Cheesiness					
Definitely too weak	13.40				
Definitely too strong	23.40	0.01	0.66	0.52	0.84
Meaty Taste					
Definitely too weak	9.80				
Definitely too strong	17.03				
Pineapple Taste					
Definitely too weak	10.30				
Definitely too strong	36.20	-0.03	-1.12	0.47	0.08
Overall Flavor Blend					
Definitely too weak	13.97				
Definitely too strong	60.50	0.05	-0.58	0.48	0.79

**significant at $\alpha = 0.05$*

For Hawaiian Pizza B (Table 8), the respondents significantly penalized the product when found to be having definitely too little amount of cheese and definitely too little amount of meat. The drop in the overall liking when not judged to be at JAR level is 2.21 for the amount of cheese and 2.44 for the amount of meat. Increasing both the amount of cheese and the amount of meat may cause an increase in the overall liking score of Hawaiian Pizza B. Penalty test is not significant when the respondents found the product to be having definitely too big toppings, definitely too much cheese, definitely too much meat, definitely too much pineapple, definitely too strong cheesiness, definitely too strong pineapple taste and definitely too strong overall flavor blend. The mean drops were not obtained for all the remaining attributes because the percentages of respondents were lower than 20%.

4.4 Comparison of ordinary and bootstrapping penalty analysis

Tables 9 and 10 show the comparison of the results of ordinary penalty analysis and bootstrapping penalty analysis for Hawaiian Pizza A and Hawaiian Pizza B respectively. The results include the percentage of respondents for the non-JAR level, mean drop and adjusted bootstrap mean drop.

Based on the results of ordinary penalty analysis (Table 9), the respondents penalized Hawaiian Pizza A when they rated the product as having definitely too little amount of meat, definitely too little amount of pineapple, definitely too weak meaty taste, definitely too strong pineapple taste and definitely too much amount of meat. When penalty analysis is combined with bootstrap method, the respondents penalized the same attributes. However, only the "definitely too little" amount of meat and "definitely too little" amount of pineapple were found to have significant effect on the overall liking score of Hawaiian Pizza A. For Hawaiian B (Table 10), the respondents penalized the product for having definitely too little amount of cheese, definitely too little amount of meat, definitely too strong cheesiness, definitely too much amount of cheese, and definitely too big size of the toppings. However, results of bootstrapping penalty analysis showed that the only significant attributes are "definitely too little" amount of cheese and "definitely too little" amount of meat.

Based on Tables 9 and 10, it can also be observed that the percentages of respondents for the non-JAR levels are almost the same for ordinary penalty analysis and bootstrapping penalty analysis. Also, it can be noted that results obtained for the adjusted bootstrap mean drops are almost twice as the value of the mean drops. This means that when penalty analysis is used together with bootstrap method, the amount of mean drops from the overall liking score is higher compared to an ordinary penalty analysis.

Table 9. Comparison of Ordinary and Bootstrapping Penalty Analysis for Hawaiian Pizza A

Attributes	% Respondents (Ordinary PA)	% Respondents (Bootstrapping PA)	Mean Drop	Adjusted Mean Drop (Bootstrap)
Size of the Toppings				
Definitely too small	3.57	5.77		
Definitely too big	17.86	16.13		
Amount of Cheese				
Definitely too little	17.86	17.00		
Definitely too much	28.57	25.90	-0.60	-1.25
Amount of Meat				
Definitely too little	32.14	32.67	1.23	2.25*
Definitely too much	21.43	20.90	0.10	0.20
Amount of Pineapple				
Definitely too little	28.57	29.10	1.18	2.38*
Definitely too much	7.14	6.67		
Saltiness				
Not salty at all	0	0		
Definitely too salty	17.86	17.50		
Cheesiness				
Definitely too weak	21.43	18.67		
Definitely too strong	25.00	27.27	-0.44	-0.86
Meaty Taste				
Definitely too weak	21.43	20.37	0.83	1.66
Definitely too strong	21.43	20.50	-0.83	-1.65
Pineapple Taste				
Definitely too weak	10.71	9.87		
Definitely too strong	28.57	30.00	0.11	0.15
Overall Flavor Blend				
Definitely too weak	14.29	11.70		
Definitely too strong	42.86	44.27	-0.39	-0.78

**significant at α = 0.05*

Table 10. Comparison of Ordinary and Bootstrapping Penalty Analysis for Hawaiian Pizza B

Attributes	% Respondents (Ordinary PA)	% Respondents (Bootstrapping PA)	Mean Drop	Adjusted Mean Drop (Bootstrap)
Size of the Toppings				
Definitely too small	10.00	10.73		
Definitely too big	36.67	36.90	0.07	0.22
Amount of Cheese				
Definitely too little	30.00	31.33	1.56	3.14*
Definitely too much	23.33	22.67	0.21	0.49
Amount of Meat				
Definitely too little	23.33	22.60	0.90	1.87*
Definitely too much	26.67	27.00	-0.03	-0.03
Amount of Pineapple				
Definitely too little	10.00	10.23		
Definitely too much	26.67	26.87	-0.22	-0.42

Saltiness Not salty at all Definitely too salty	 13.33 13.33	 13.47 12.57		
Cheesiness Definitely too weak Definitely too strong	 13.33 23.33	 13.40 23.40	 0.33	 0.66
Meaty Taste Definitely too weak Definitely too strong	 10.00 16.67	 9.80 17.03		
Pineapple Taste Definitely too weak Definitely too strong	 10.00 36.67	 10.30 36.20	 -0.57	 -1.12
Overall Flavor Blend Definitely too weak Definitely too strong	 13.33 60.00	 13.97 60.50	 -0.26	 -0.58

**significant at α = 0.05*

4.5 Implications on product development

Penalty analysis combined with bootstrap method is a great way to allow the product developers decide as to what sensory attributes should be improved first in order to increase overall liking score.

Based on the results of bootstrapping penalty test, Hawaiian Pizza A was significantly penalized by the respondents for having too little amount of meat and too little amount of pineapple. Since the mean drop is larger for the amount of pineapple, the product developer may increase this first and check if the overall liking score will improve. The developer may also increase the amount of both the meat and pineapple simultaneously. Hawaiian Pizza B, on the other hand, was significantly penalized for having too little amount of cheese and too little amount of meat. Increasing the amount of these ingredients may improve the overall liking score of the product. The developer may also decide to prioritize increasing the amount of cheese first since it has a larger mean drop compared to the amount of meat. However, it should be noted that increasing the amount of any of these ingredients would entail additional cost on the part of the owner.

Adjusting attributes that significantly penalized Hawaiian Pizza A and Hawaiian B may contribute to an increase in overall liking score. However, this is not a guarantee that the hedonic overall liking score will increase by the same amount of mean drops when the attributes not in the JAR levels are corrected. They simply suggest what attributes should be prioritized when improving the product (Rothman, 2007). Caution must be likewise taken as the change in one attribute can cause a significant change in the other attributes.

To determine which of the two pizza products should be improved, commercialized and launched, the results of the overall acceptability percentage, preference test and bootstrapping penalty analysis were compared. Since Hawaiian Pizza B obtained a higher overall acceptability rating and is not significantly preferred over Hawaiian Pizza A, the product developer may focus first on improving attributes that significantly penalized Hawaiian Pizza B.

5. Conclusions and Recommendations

Ordinary penalty analysis and bootstrapping penalty analysis were done on the sensory evaluation data of Hawaiian Pizza A and Hawaiian Pizza B. The respondents penalized Hawaiian Pizza A for having too much amount of cheese, too much of meat, too much pineapple, too strong cheesiness, too strong pineapple taste and too strong overall flavor blend. However, results of bootstrapping penalty analysis showed that only the "too little amount of meat" and "too little amount of pineapple" have significant mean drops on the overall liking. On the other hand, Hawaiian Pizza B was penalized for having too big size of the toppings, too little amount of cheese, too much amount of cheese, too little amount of meat, too much amount of meat, too much amount of pineapple, too strong cheesiness, too strong pineapple taste and too strong overall flavor blend. The only significantly penalized attributes are the "too little amount of cheese" and "too little amount of meat."

Based on the results of this study, it can be concluded that bootstrap test could be used together with penalty analysis to allow statistical testing of the mean drops. It is then recommended to adjust the level of the attributes that were significantly penalized by the respondents and to conduct another in-house sensory evaluation. Bootstrapping penalty analysis can be applied again to determine if there would be an improvement in the overall liking score and if the mean drops would be corrected. It is also recommended to focus on improving Hawaiian Pizza B over Hawaiian Pizza A since it obtained a higher overall acceptability percentage.

References

AMERINE, M., PANGBORN, R. and ROESSLER, E., 1965, *Principles of Sensory Evaluation of Food*, New York: Academic Press.

ANON, 2003, Triangle plots: Graphic display of "just right" scale data, *Research on Research* 56: 1-6.

BAOSHENG, G., 2005, *Product Testing: Handle with Care*, China: IPSOS Insight.

CHAMBERS, E., IV and WOLF, M., 1996, Sensory testing methods, *ASTM Manual Series*: MNL 26,West Conshohoken, PA: ASTM International.

CHERNICK, M., 1999, *Bootstrap Methods: A Practitioner Guide*, New York: John Wiley & Sons, Inc.

DAVIDSON, A., HINKLEY, D. and SCHECHTMAN, E., 1986, Efficient bootstrap simulation, *Biometrika* 73(3): 555-566.

DIACONIS, P. and EFRON, B., 1983, Computer-intensive methods in statistics, *Scientific American* 248 (5): 116-130.

EFRON, B. and GONG, G., 1983, A leisurely look at the bootstrap: The jackknife and cross validation, *The American Statistician* 37: 36-48.

EFRON, B. and TIBSHIRANI, R., 1993, *An Introduction to the Bootstrap*, New York: Chapman and Hall, Inc.

FAN, X., 1994, *Does Bootstrap Procedure provide Biased Estimates? An Empirical Examination for a Case of Multiple Regression*, East Lansing, MI: National Center for Research on Teacher Learning.

GATCHALIAN, M., 1989, *Sensory Evaluation Methods for Quality Assessment and Development*, University of the Philippines, Quezon City.

GATCHALIAN, M. and BRANNAN, G., 2009, *Sensory Quality Measurement: Statistical Analysis of Human Responses*, Quezon City, Philippines: Quality Partners Company, Ltd.

JONES, L., PERYAM, D. and THURSTONE, L., 1955, Development of a scale for measuring soldiers' food preferences, *Food Research* 20: 512-520.

KEMP, S., HOLLOWOOD, T. and HORT, J., 2009, *Sensory Evaluation: A Practical Handbook*, West Sussex, UK: John Wiley & Sons, Ltd., Publication.

LAWLESS, H. and HEYMANN, H., 1999, *Sensory Evaluation of Food: Principles and Practices*, Gaithersburg, Maryland: Aspen Publishers, Inc.

LEVINE, D. and STEPHAN, D., 2010, *Even You Can Learn Statistics: A Guide for Everyone who has ever been Afraid of Statistics*, Upper Saddle. River, NJ: Pearson Education, Inc.

MEILGAARD, M., CIVILLE, G. and CARR, B., 1999, *Sensory Evaluation Techniques*, 2nd ed., Boca Raton, FL: CRC Press.

MEULLENET, J., XIONG, R. and FINDLAY, C., 2007, *Multivariate and Probabilistic Analyses of Sensory Science Problems*, 1st ed., Iowa, USA: Blackwell Publishing.

MASON, R. and NOTTINGHAM, S., 2002, *Sensory Evaluation Manual*, Workshop at Naresuan University, Phitsanulok, Thailand, Available at: http://www.scribd.com/doc/8940001/Sensory-Evaluation-Manual#scribd

PACZKOWSKI, W., 2009, *Technical Memorandum on Penalty Analysis*. Data Analytics Corp.

PERYAM, D.R. and PILGRIM, P.J., 1957, Hedonic scale method for measuring food preferences, *Food Technology* 11:9–14.

PLAEHN, D., 2009, *Understanding Penalty Analysis*, Insights Now, Inc.

PLAEHN, D. and HORNE, J., 2008, A regression-based approach for testing significance of "just-about-right" variable penalties, *Food Quality and Preference* 19: 21-32.

PRELL, P., 1976, Preparation of reports and manuscripts which include sensory evaluation data, *Food Technology*.

ROTHMAN L., 2007, *The Use of Just-About-Right (JAR) Scales in Food Product Development and Reformulation in Consumer-Led Food Product Development*, Boston: CRC Press.

SCHMIDHEINY, K., 2010, *The Bootstrap: Short Guides to Microeconometrics*,Unversi tatPompeuFabra.

STINE, R., 1989, An introduction to bootstrap methods, *Sociological Methods and Research* 18:243-291.

STONE, H. and SIDEL, J., 2004, *Sensory Evaluation Practices*, 3rd ed., London/New York: Academic Press/Elsevier.

TIBSHIRANI, R., 1985, *How Many Bootstraps*? Department of Statistics, Stanford University Technical Support.Standford, California.

WHISTON S., 2009, *Principles and Applications of Assessment in Counseling*. 3rd ed., Belmont, CA: Brooks/Cole, Cengage Learning.

XIONG, R., MEULLENET, J. and FINDLAY, C., 2007, *Multivariate and Probabilistic Analyses of Sensory Science Problems*, Blackwell Publishing, 207-212.

XIONG, R. and MEULLENET, J., 2009, Bootstrapping penalty analysis, *ASTM International* 63-66.