Food Sci. Biotechnol. Vol. 15, No. 4, pp. 494 ~ 498 (2006)

MINI REVIEW

Food Science and Biotechnology

© The Korean Society of Food Science and Technology

Sensory Difference Testing: The Problem of Overdispersion and the Use of Beta Binomial Statistical Analysis

Hye-Seong Lee[1,2,3]* **and Michael O'Mahony**[1]

[1]*Department of Food Science and Technology, University of California, Davis, CA 95616, USA*
[2]*Based at Unilever Food & Health Research Institute Vlaardingen, Olivier van Noortlaan 120, 3133 AT Vlaardingen, The Netherlands*
[3]*Department of Food Science and Technology, Ewha Womans University, Seoul 120-750, Korea*

Abstract An increase in variance (overdispersion) can occur when a binomial statistical analysis is applied to sensory difference test data in which replicate sensory evaluations (tastings) and multiple evaluators (judges) are combined to increase the sample size. Such a practice can cause extensive Type I errors, leading to serious misinterpretations of the data, especially when traditional simple binomial analysis is applied. Alternatively, the use of beta binomial analysis will circumvent the problem of overdispersion. This brief review discusses the uses and computation methodology of beta binomial analysis and in practice evidence for the occurrence of overdispersion.

Keywords: sensory evaluation, sensory difference tests, binomial statistics, overdispersion, beta binomial test, gamma

Introduction

Sensory difference tests are used in food science to determine if judges can discriminate between two food stimuli which are so similar that they can be described as confusable. Such tests are used for quality assurance, ingredient specifications, product development, and studies of the effects of process changes, packaging changes, and product storage. The tests require judges to demonstrate their ability to distinguish between the two foods stimuli in question. Four different test methods are commonly used as outlined: a 2-Alternative Forced Choice (2-AFC) method, sometimes called the paired comparison where the judge is presented with two stimuli and is required to select one of the stimuli (e.g., the sweeter, fruitier, firmer, etc.) to prove an ability to distinguish between the two; a 3-Alternative Forced Choice (3-AFC) method that is similar, except that the specified target stimulus is one of three, e.g., a judge might be told to select from three beverages the one which is sweeter while the other two are less sweet and are identical; the triangle test which is similar to 3-AFC wherein the judge is not told the nature of the difference but rather that two stimuli are identical while one is different and the test is to select the different one; and finally the duo-trio test that requires a judge to taste a stimulus designated as the 'standard' or 'reference' after which a pair of stimuli are to be discriminated, one being the same as the 'standard' while the other is different. The test is to indicate which stimulus is the same as the 'standard'. Each of the tests outlined are generally repeated to ensure that judges are performing better than at chance levels. Traditionally, the results are analyzed statistically for significance using procedures that are based on binomial distributions, although recently, Thurstonian analyses for determining the extent of the differences have been introduced (1). This article will consider some of the problems that are associated with binomial analysis for establishing significance and suggest a means to avoid such problems.

Binomial statistics

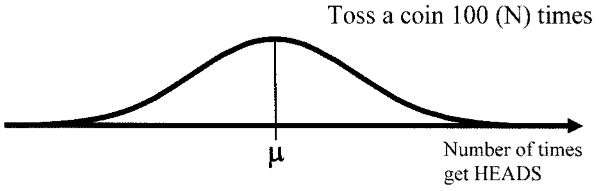
It is traditional that statistical tests based on the binomial distribution are used to determine whether the proportion of difference tests that are performed correctly is greater than chance, thus concluding that the difference was 'significant'. Yet, such binomial statistics were designed to analyze the results of tossing coins or dice when the probability of getting a target result ('heads' or 'six') is constant for each item (1/2 or 1/6). Coins and dice can be considered as clones and any variability in the expected proportion of 'heads' or 'sixes' that is obtained, is due to chance and described by the binomial distribution.

For example, if a coin is tossed 100 (N) times and the probability of getting 'heads' (p) is 1/2, and the probability of getting 'tails' (q) is also 1/2, then the expectation is that 'heads' (and 'tails') will occur 50 times. That result, however, will not always occur. If the coin is tossed 100 times, 'heads' will sometimes occur more often and sometimes less often. This variation in expectations is described by the binomial distribution which has a mean of 50 (Np, the most frequent occurrence) and a variance given by the formula Npq = 25. See Fig. 1.

The problem of overdispersion

With a set of difference tests performed by a single judge during a single experi-mental session, it could be argued that the probability of getting a target result (test correct) is constant over repeated tests, should there not be fatigue or adaptation effects over replicate testings. If data from several judges were to be combined, however, this situation would no longer hold because each judge can be expected to have different sensitivities and thus their probabilities of getting the target result (test correct) would vary. The assumptions for the binomial test, therefore,

*Corresponding author: Tel: 82-2-3277-3624; Fax: 82-2-3277-2862
E-mail: hyeseonglee@yahoo.com
Received November 10, 2005; accepted June 12, 2006



p = probability of getting target result HEADS = ½

q = probability of not getting target result NOT HEADS = ½

MEAN $\mu = Np = 100 \times ½ = 50$

VARIANCE $\sigma^2 = Npq = 25$

Fig. 1. The binomial distribution representing the proportion of times a coin would land as 'heads' after a total of 100 replicate tosses.

would be violated. This violation can result in Type I errors and the declaration of a 'significant' difference when the results were actually due to chance or by guessing.

The variation in the sensitivity of the judges provides an extra source of variance in the computation, more variance than would be expected from mere binomial analysis. The problem of this extra variance is termed "the problem of overdispersion".

The beta binomial distribution and gamma

There are various solutions to the problem of overdispersion (2, 3). The approach discussed in this review uses beta distributions to describe the distributions of judge sensitivities that are encountered during difference testing. The beta distributions are combined with the regular binomial distributions to give what are called beta-binomial distributions. These combinations are the basis for the beta-binomial statistical analysis for difference tests. The beta-binomial with the extra variance brought in by the beta distribution will have greater variance than the binomial distribution alone. Greater variance means greater difficulty in rejecting the null hypothesis and declaring a 'significant' difference. In this way, it can be seen that there is a risk of Type I error if binomial analysis is used, rather than a beta-binomial analysis.

In an early study, Harries and Smith (4) investigated the beta-binomial analysis for triangle tests and illustrated some of the beta distributions. One illustration indicated a small range of judge sensitivity while a second indicated more of a bimodal distribution. More recently, the beta-binomial test has been further developed by Ennis, Bi and their coworkers (5-8) who have described overdispersion by a gamma index (γ). In the gamma index, a gamma value of zero indicates no overdispersion while a gamma value of unity indicates maximum overdispersion. In general, gamma values are intermediate between zero and unity (see examples later).

It is instructive to consider the shapes of some beta distributions as shown in Fig. 2. At the top of Fig. 2(a) is represented by a single line, giving the constant probability of getting a target result (1/2, coins; 1/6 dice); gamma would be zero. Coins and dice, however, can be considered as clones and should human clones be the ones testing, there would, once again, be a single probability. However, human judges are not clones and so there will be a distribution of probabilities (the beta distribution), according to the sensitivity of the judges. There might be both sensitive and insensitive judges, and therefore, a bimodal distribution (Fig. 2(b), continuous line) would result; gamma would not be zero. Alternatively, the judges may have very similar sensitivity and produce a tightly bunched distribution (see Fig. 2(b), dashed line).

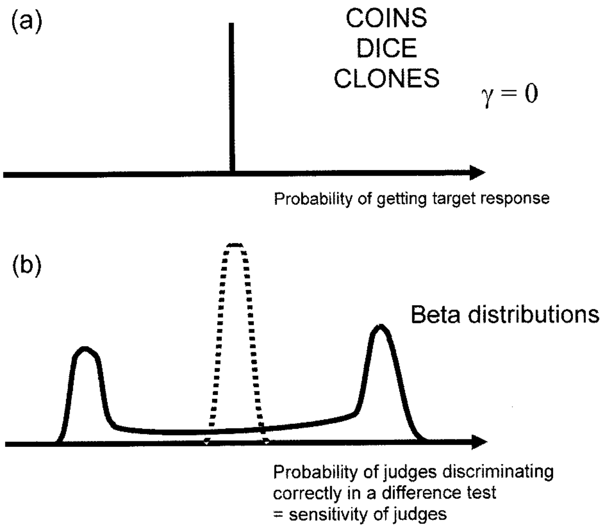

(a)
COINS
DICE
CLONES
γ = 0
Probability of getting target response
(b)
Beta distributions
Probability of judges discriminating correctly in a difference test = sensitivity of judges

Fig. 2. Beta distributions. (a) represents a fixed probability value for getting a target result when tossing coins or dice, etc. (e.g., p= 1/2 or 1/6); (b) represents possible distributions of the sensitivity of groups of judges. The continuous line indicates two groups of judges, one of which is sensitive and the other insensitive. The dotted line indicates a single group of judges with a narrow range of sensitivity.

When gamma is zero, there is no overdispersion and the regular binomial statistical analysis method can be used. Thus, analysis is a simple matter of looking up values in tables in standard statistical texts (for example, pages 408-413 in reference 9). When gamma is greater than zero, indicating overdispersion, the analysis must be modified. The beta distribution is combined with the binomial distribution to yield the beta binomial distribution which is then used as the basis for the analysis. In practice the analysis is relatively simple. Once a gamma value has been calculated, the method becomes a matter of using alternative tables of values. Such tables as published by Bi and Ennis (10) indicate the proportion of tests that are required to be correct in declaring a 'significant' difference for various values of gamma. For example, an international beverage company wished to determine whether an ingredient reduction could be detected by their trained sensory panel of 10 judges. The company required each judge to perform 10 one-tailed 2-AFC tests, and to increase the statistical power of the sensory method, they pooled the results for the 10 judges to give a total of 100 tests. Had gamma been zero, indicating no overdispersion, a simple binomial analysis would have been appropriate for significance testing and the number of tests required to be performed correctly to enable a declaration of a significant difference ($p<0.05$) that would have been 59. A subsequent analysis, however, indicated that there had been overdispersion; because judges varied in their sensitivities and the gamma value was not zero, but 0.5. Accordingly, the number of tests required to be performed correctly to declare a 'significant' difference was 70 (i.e., greater than 59). The experimenters, however, were not aware of the presence and likely complication of overdispersion and used simple binomial statistical analysis. The number of tests the judges performed correctly was 65. Accordingly, because this value exceeded 59, the experimenters went on to report a significant difference between the original and the reformulated beverages. However, 65 tests did not exceed the actual required value of 70 and, therefore, a Type I error was committed. The experimenters had incorrectly declared a significant difference and concluded that the ingredient reduction was not successful.

Computation of gamma and its significance

Before using the statistical tables of gamma, it is important to be able to determine the gamma value. This value can be computed using the formula below.

$$\gamma=\frac{n_R S}{\bar{p}(1-\bar{p})N_J(n_R-1)}-\frac{1}{(n_R-1)} \qquad (1)$$

n_R= number of replicates per judge
N_J= number of judges
$\bar{p}$ = mean probability value

$$= \frac{\text{total number of correct tests}}{\text{total number of tests}} = \frac{X}{N_J n_R}$$

$$S = \sum\left(\frac{X}{n_R}-\bar{p}\right)^2$$

$\sum$: summed over people

$\left(\frac{X}{n_R}-\bar{p}\right)^2$: deviation of each person from mean

The term 'S' in the formula above expresses the variability of the judges' sensitivity.

To test whether gamma is significantly greater than zero, Tarone's z-statistic is used and is computed using the formula below.

$$Z=\frac{(E-N_J n_R)}{\sqrt{2N_J n_R(n_R-1)}} \qquad (2)$$

Z = number of standard deviations from mean of normal distribution
X = number of correct tests for each person

$\bar{p}$ = mean probability value =

$$= \frac{\text{total number of correct tests}}{\text{total number of tests}} = \frac{\sum X}{N_J n_R}$$

$n_R\bar{p}$ = average number of correct tests per person

$$E=\frac{\sum(X-n_R\bar{p})^2}{\bar{p}(1-\bar{p})} \qquad (3)$$

It can be seen that the variability between judges is expressed by the term 'E'. The calculated z-value can be used with normal distribution tables (for example, pages 404-405 in reference 9) to determine the probability of obtaining such a gamma value by chance. Software for these computations is available at IFPrograms, Institute for Perception, Richmond, VA, USA.

The beta-binomial distribution has a simple relationship to the binomial distribution. The mean of the beta-binomial distribution is $N\bar{p}$, where $\bar{p}$ is the mean probability of getting a test correct (if 10 judges each performed 10 tests and the total number correct was 68, $\bar{p}$ = 68/100). (In the same way, $\bar{q}$ is the mean probability of getting a test incorrect). The variance ($N\overline{pq}$) is modified by a simple multiplier: $[1 + (n_R - 1)\gamma]$. For experiments in which the data analysis is in terms of d', the beta-binomial approach fits conveniently into a Thurstonian framework (5). The required increase in variance for d' is obtained using the same multiplier.

When does overdispersion occur?

In practice, over-dispersion (a γ value significantly greater than zero) is sometimes encountered, but not always. The question of whether different test protocols are more prone to overdispersion than others has been asked. There is a good reason for asking this. Judges use different cognitive strategies for different tests (1), strategies that might elicit a routine which could render judges to perform in similar (similar sensitivity, low gamma) or dissimilar ways (different sensitivities, high gamma).

Rousseau and O'Mahony (working with orange drinks) found overdispersion with triangle tests in one study (11) but not in another (12). Braun *et al.* (13) using a model system of NaCl solutions and 2-AFC tests, reported overdispersion, while recent unpublished work in our laboratory, using the same method and stimuli, reported overdispersion in one study and no overdispersion in another. The same study compared 2-AFC, 3-AFC, triangle and duo-trio tests and indicated overdispersion for the triangle tests in one study and the 2-AFC tests in another.

In another study, Delwiche and Ligget (14) required judges to perform paired preference and 2-AFC tests for fruit flavored beverages, chips, and cookies. The occurrence of significant gamma values was inconsistent. For both tests, gamma values occurred for two out of five studies and did not occur for the same foods. Increasing the number of replicate tastings also had little effect. These investigators also compared the performance of judges on 2-AFC, 3-AFC, triangle and duo-trio tests with cherry flavored drinks; a significant gamma value was obtained only with the duo-trio tests, unlike recent study results in our laboratory. Again using cherry flavored drinks and 2-AFC tests; the performance of judges over a 10 day period was measured. Performance continued to be inconsistent; significant overdispersion occurred on three of the days of testing. In light of these studies, there does not appear to be any particular pattern to support the notion that one test may be more prone to overdispersion than another, yet even this evidence is sparse.

To approach the problem differently, it is worth reconsidering how values of gamma should be interpreted. Since that judges are not human clones, it may be asked why cases occur when gamma values are not significant and there is no significant overdispersion. This finding would only be possible if the addition of a beta distribution to a binomial distribution had minimal effects on the shape and variance of the latter. As Harries and Smith (4) indicated, a beta distribution that was fairly compact and clustered around the mean of a binomial distribution would have just such a minimal effect. This result would mean that the sensitivities of the judges in the discrimination tests were close to the mean. As Harries and Smith (4) further remarked, a beta distribution that was more scattered or even bimodal would have a substantial effect on the binomial distribution and thus produce a significant gamma value. Bimodal distributions can occur in preference testing when the judges are split on their preferences, and also with difference tests if the group of judges, who happen to be sampled, includes one group that is sensitive and another that is insensitive.

To support these considerations, we performed experiments in our laboratory using a sample of judges made up of a group of knowingly less sensitive and a group of more sensitive (15). Within each group of judges, the sensitivities were fairly uniform and the gamma values for the 2-AFC and 3-AFC tests were zero. When the two groups were combined, however, the sensitivities were no longer uniform but rather were bimodal. Accordingly, significant gamma values (0.04, 0.06), indicating overdispersion, were obtained. In a second experiment, the less sensitive and more sensitive judges performed 2-AFC tests, and once again, within each group, zero gamma values were obtained while a significant gamma value (0.07) was resulted when the two groups were combined. When the sensitivity of the less sensitive group was increased, however, by using a warm-up procedure (16-18), the combination of two groups who were now comparable in their sensitivity resulted in a non-significant gamma value, indicating the absence of overdispersion. Thus, the occurrence of overdispersion appears to be more a result of chance sampling than of any particular measurement protocol.

Practical applications of beta-binomial statistics

The beta binomial allows the combination of consumers and replicate testings in the data, to increase the sample size. This approach might be acceptable in psychophysics, where a large sample size might be necessary for such tasks as fitting a response curve (Receiver Operating Characteristic, ROC curve) for signal detection analysis of a small group of judges (19). Such an analysis is used to determine the decision rules or cognitive strategies used by judges in discrimination tests. The technique, however, should be used with caution when sampling consumers from the population for their discrimination ability or preferences. For example, if 10 consumers each performed 5 difference tests, even though the sample size can be treated as 50, it remains that only 10 actual consumers will have participated and is hardly a representative sample.

It is also recognized that gamma values can only be calculated to reveal possible overdispersion effects when judges perform replicate tests. If each judge were only to perform a single test, it would not be possible to determine the existence of overdispersion. Thus, it is recommended that judges should repeat the testing process at least once, to enable the presence of overdispersion to be detected and importantly, to avoid the possibility of Type I errors. Overdispersion can be likened to interaction in the analysis of variance; its presence can only be detected with replication.

Conclusions

For significance testing with sensory difference tests, a binomial statistical analysis is generally used. To increase the power of the test, experimenters sometimes increase the sample size by combining judges and their replicate tastings. However, this practice ignores the problem of overdispersion and can result in a Type I error (wrongly declaring a significant difference). The reason for the overdispersion is that judges vary in their sensitivity and that by combining their data, the variance is increased beyond that described by binomial statistics. This extra variance, however, termed overdispersion, can be accounted for by a beta distribution. Incorporating beta distribution into the statistical analysis yields beta-binomial analysis which avoids the problem of overdispersion and likelihood of Type I errors. To detect problems of overdispersion, it is necessary for each judge to perform more than one test. The performance of replicate tests per judge should be routine. There are further cautions, however. The beta-binomial deals with sensitivity variation due to judge heterogeneity and assumes that sensitivity is constant for a given judge over replicate tests. This assumption may not hold for all product testing situations. If sensitivity varied over replicate tests, the statistical analysis would require an even more complex beta-beta binomial analysis. Thus, even though replication is required to detect the presence of overdispersion, the number of replicates chosen should be approached with caution since too many replicate tests could produce 'taste fatigue' which in turn, would reduce the sensitivity of the judge over replicate tests, thus requiring a more complex beta-beta binomial analysis. Preliminary experimentation will always be necessary for determining the appropriate number of replicate tastings per session to provide adequate testing and not inducing taste fatigue.

References

1. Lee H-S, O'Mahony M. Sensory difference testing: Thurstonian models. Food Sci. Biotechnol. 13: 841-847 (2004)
2. Brockhoff PB. The statistical power of replications in difference tests. Food Qual. Prefer. 14: 405-417 (2003)
3. Brockhoff PB, Schlich P. Handling replications in discrimination tests. Food Qual. Prefer. 9: 303-312 (1998)
4. Harries JM, Smith GL. The two-factor triangle test. J. Food Technol. 17: 153-162 (1982)
5. Bi J, Ennis DM. A Thurstonian variant of the beta-binomial model for replicated difference tests. J. Sens. Stud. 13: 461-466 (1998)

6. Bi J, Ennis DM. The power of sensory discrimination methods used in replicated difference and preference tests. J. Sens. Stud. 14: 289-302 (1999a)
7. Bi J, Templeton-Janik L, Ennis JM, Ennis DM. Replicated difference and preference tests: how to account for inter-trial variation. Food Qual. Prefer. 11: 269-273 (2000)
8. Ennis DM, Bi J. The beta-binomial model: accounting for inter-trial variation in replicated difference and preference tests. J. Sens. Stud. 13: 389-412 (1998)
9. O'Mahony M. Sensory Evaluation of Food. Statistical Methods and Procedures. Marcel Dekker, Inc., NewYork, NY, USA. pp.408-413 (1986)
10. Bi J, Ennis DM. Beta-binomial tables for replicated difference and preference tests. J. Sens. Stud. 14: 347-368 (1999b)
11. Rousseau B, O'Mahony M. Investigation of the effect within-trial retasting and comparison of the dual-pair, same-different, and triangle paradigms. Food Qual. Prefer. 11: 457-464 (2000)
12. Rousseau B, O'Mahony M. Investigation of the dual pair method as a possible alternative to the triangle and same-different tests. J. Sens. Stud. 16: 161-178 (2001)
13. Braun V, Rogeaux M, Schneid N, O'Mahony M, Rousseau B. Corroborating the 2-AFC and 2-AC Thurstonian models using both a model system and sparkling water. Food Qual. Prefer. 15: 501-507 (2004)
14. Delwiche J, Ligget R. Accounting for between-subject variance in discrimination and preference tasks. J. Sens. Stud. 20: 48-61 (2005)
15. Angulo O, Lee HS, O'Mahony M. Sensory difference tests: Over-dispersion and warm-up. Food Qual. Prefer. 17: in press (2006)
16. Dacremont C, Sauvageot F, Duyen TH. Effect of assessors expertise on efficiency of warm-up for triangle tests. J. Sens. Stud. 15: 151-152 (2000)
17. O'Mahony M, Thieme U, Goldstein LR. The warm-up effect as a measure of increasing the discriminability of sensory difference tests. J. Food Sci. 53: 1848-1850 (1988)
18. Thieme U, O'Mahony M. Modifications to sensory difference test protocols: the warmed up paired comparison, the single standard duo-trio and the A-not A test modified for response bias. J. Sens. Stud. 5: 159-176 (1990)
19. Hautus MJ, Irwin RJ. Two models for estimating the discriminability of foods and beverages. J. Sens. Stud. 10: 203-215 (1995)