The Institute for Perception

For technical reports from The Institute for Perception or for information about short courses, please visit www.ifpress.com or email us at mail@ifpress.com

Reprinted from IFPress (2018) 21(2) 3,4

Consumer Relevant Confidence in Sensory Difference Tests

John M. Ennis and Daniel M. Ennis

Introduction: Sensory difference testing is a business relevant topic, as evidenced by its use in cost reductions, ingredient and process changes, compliance with health initiatives, and quality control. In recent years there has been an increasing emphasis on measuring the size of possible inter-product differences and making confidence statements about them, as opposed to conducting hypothesis tests to detect possible differences[1,2]. This is a valuable shift that has important implications for how we view difference testing for the objectives listed above. In current practice, sensory difference testing continues to rely on statistical techniques that in many ways fail to correspond to the issues difference tests are conducted to address. In this report, we discuss a more appropriate statistical framework to address more directly the objectives of difference tests and the variability in the measurements upon which the framework rests.

Scenario: You are responsible for sensory testing in a confectionery company and you are tasked with considering an ingredient change to one of your prominent candy bar products. One of your objectives is to ensure that there is a low likelihood that your consumers will notice the change. In a project like this, you would normally conduct a difference test and base your recommendations on whether your discrimination panel can detect a difference, as is standard practice in many sensory programs. You design a difference test to determine whether to proceed with the ingredient change. Recent research on the relative power of difference testing methods leads you to base your testing on the Tetrad test which has been shown to be more powerful than the Triangle test[3]. You have a panel size of 30 and, following your standard approach, you conduct a replicated Tetrad test in which each panelist completes 3 trials for a total of 90 evaluations. Using the beta-binomial model, you find no evidence for panelist heterogeneity[4] so you combine your data and obtain 39 correct responses out of 90.

The p-value associated with a standard binomial analysis of these data is 0.03, as shown in Figure 1, indicating that you would normally reject the ingredient change. However, you also now consider the size of the difference using a Thurstonian model for the Tetrad test[5] and find that $d' = 0.768$ with an estimate variance of 0.048.

The value of d' is an estimate of the size of the sensory difference, δ, which is scaled in units of the common perceptual standard deviation of each product. You have learned that a Just Noticeable Difference (JND) of 1.0 corresponds very closely to a δ of 1.0 and is sometimes used as a consumer relevant action standard[6,7]. You wonder if your sensory difference, while perhaps greater than zero, might not still be less than this action standard or an action standard that you might determine experimentally.

Hypothesis Testing Limitations: When sensory difference tests are typically conducted, results are compared to what would be expected if there was no difference between the samples. In analyzing the scenario data, there were 39 correct responses out of 90. Figure 1 shows that such a result is unlikely ($p < 0.05$) if the samples are identical.

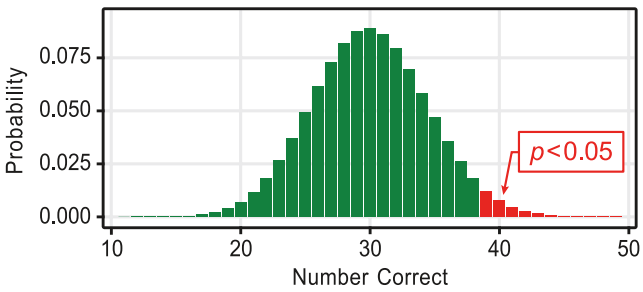

Figure 1. Probability of various outcomes in a sample of 90 independent Tetrad trials.

There are multiple critiques of the standard approach to statistical testing of difference tests. Several standard methods, including the Triangle test, require very large sample sizes to detect moderately-sized sensory differences reliably[8]. This critique argues that sample sizes should be chosen by considering both the testing method and the size of the difference one wishes to detect. Fortunately, your adequately powered Tetrad test addresses this critique. There is a second critique, however. Without considering the size of the difference that matters to consumers, it is not possible to know what size difference one wishes to detect. This critique echoes your second concern about why a δ of 1 should receive special consideration; if you choose a consumer-relevant action standard, there remains the difficulty of how to interpret a significant difference.

Interpreting Difference Testing Results: Even when power and consumer relevance are considered, there is still a need to know how to interpret significantly different results. For example, suppose one wishes to detect a δ of 1 in a Tetrad test. In this case, 65 evaluations are required for 80% power. If one desires higher power, say 95%, 110 evaluations are required. But the challenge in both cases is that effect sizes between 0.5 and 0.75 are frequently rejected, as Figure 2 demonstrates.

Consumer Relevant Confidence: To build a statistical test appropriate to the business relevant questions that sensory difference tests seek to answer, we must answer the question: "How confident are we that the underlying difference is not consumer relevant?" If τ_R is the size of the consumer relevant difference, the value of which we only have an estimate, we can formulate the question by finding the probability that our measure of the sensory difference is less than our measure of the consumer relevant difference. We call this probability *consumer relevant confidence*. Given an estimate of τ_R, we propose that sensory difference tests should be designed so that there is high consumer relevant confidence and, following a sensory difference test, one should compute one's confidence that an underlying sensory difference is not consumer relevant. Thus, consumer relevant confidence is informative regardless of whether the test result is statistically significant.

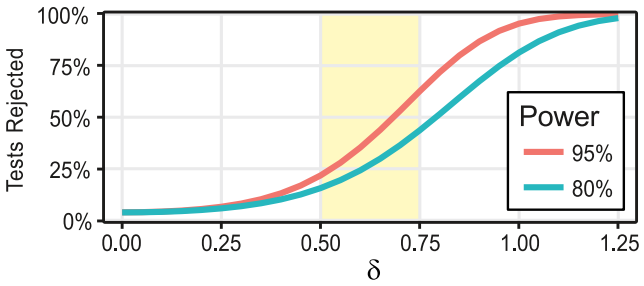

Figure 2. The probability of rejecting product changes at 95% or 80% power when the actual δ varies.

Accounting for Variability: To compute consumer relevant confidence it is necessary to account for the fact that δ is not known with certainty and that τ_R is not known with certainty. Thus, we need a solution that accounts for variability in the estimates of both δ and τ_R. Fortunately, this solution is provided by recently developed theory for multiplicative comparisons[9]. Given two normal random variables ***X*** and ***Y*** corresponding to a measure of the action standard and a measure of the detected difference, the theory addresses the question:

$$\text{What is } P(X>Y \text{ and } X>0)?$$

This theory can be used to compute the probability that the consumer relevant action standard is greater than the sensory difference measure, which means this theory can be used to compute consumer relevant confidence. Suppose that: a) $\tau_R = 1$, b) an experiment with 100 same and 100 different pairs in a same/different test provides a measure of the variance in the consumer relevant action standard[7], and c) Tetrad testing occurs for various values of δ. Figure 3 shows how expected consumer relevant confidence compares with the percentage of tests rejected. The yellow highlighted region shows cases where a change has a high probability of rejection using hypothesis testing, but where consumer relevant confidence remains high. Note that higher power only exacerbates this issue.

Applying the Theory to the Ingredient Change Problem: Based on the theory described above, you conduct a Same-Different test with 200 consumers, 100 of each type of pair, to assess the point at which consumers begin to call samples "different". In your experiment, each respondent evaluates either a "same" or "different" pair as shown in Tables 1 and 2.

Response	Same Pair	Different Pair
"Same"	54	50
"Different"	46	50

Table 1. Same-Different results.

Parameter	Estimate	Variance in Estimate
δ	0.608	0.300
τ_R	1.040	0.013

Table 2. Analysis of the data in Table 1.

Based on prior research you know that your experienced panel data typically yields d' values that are 25% larger than the d' values your consumer data yields. These results are consistent with the fact that members of your expert panel are more sensitive than are your consumers. This increased sensitivity is also seen in your present data - the expert panel Tetrad test returned a d' value of 0.768, while your consumer Same-Different test returned a smaller d' value of 0.608.

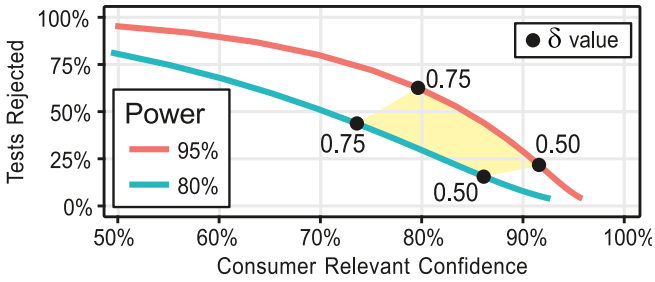

Figure 3. The yellow highlighted region shows cases where a rejection would occur, although there is high confidence that the change is not consumer relevant.

Knowing this relationship, you rescale your consumer τ_R value and variance, and compute a consumer relevant confidence level. This value is 98%. You are now quite sure that the difference is not consumer relevant and recommend the ingredient change, even though the ingredient change would have been rejected according to the standard approach.

Conclusion: Using the theory of consumer relevant confidence, we can design tests and analyze test results according to research goals rather than relying on *p*-values from difference tests. Specifically, we can compute our confidence that the sensory effect size is less than an experimentally determined consumer relevant action standard regardless of the statistical significance of the test outcome. Once consumer relevant confidence is computed, we are in a position to make an informed decision.

References

1. Cumming, G. (2014). The new statistics: Why and how. *Psychological Science,* **25**(1), 7-29.
2. Kruschke, J. K. and Liddell, T. M. (2018). The Bayesian new statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychonomic Bulletin & Review,* **25**(1), 178-206.
3. Ennis, J. M. and Jesionka, V. (2011). The power of sensory discrimination methods revisited. *Journal of Sensory Studies,* **26**(5), 371-382.
4. Ennis, D. M. and Bi, J. (1998). The beta-binomial model: Accounting for inter-trial variation in replicated difference and preference tests. *Journal of Sensory Studies,* **13**(4), 389-412.
5. Ennis, J. M., Ennis, D. M., Yip, D., and O'Mahony, M. (1998). Thurstonian models for variants of the method of tetrads. *British Journal of Mathematical and Statistical Psychology,* **51**(2), 205-215.
6. Rousseau, B. and Ishii, R. (2016). How do perceived sensory differences and preferences relate? In *The 2nd Asian Sensory and Consumer Research Symposium* in Shanghai, China.
7. Rousseau, B. (2015). Sensory discrimination testing and consumer relevance. *Food Quality and Preference,* **43,** 122-125.
8. Ennis, D. M. (1990). The relative power of difference testing methods in sensory evaluation. *Food Technology,* **44**(4), 114, 116-117.
9. Ennis, J. M. and Ennis, D. M. (2011). Confidence bounds for multi plicative comparisons. *Communications in Statistics-Theory and Methods,* **40**(17), 3049-3054.