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 (2014) 17(2) 3,4

Confidence Intervals and Consumer Relevance

Daniel M. Ennis, Benoît Rousseau, and John M. Ennis

Background: Difference testing methods such as the Triangle, Duo-Trio, Tetrad, and 2-Alternative Forced Choice (2–AFC) generate choice counts or numbers of correct responses that are often analyzed statistically as binomial variables. This analysis approach provides the basis for hypothesis tests for difference testing methods and for commonly available tables for that purpose[1]. Since any difference can be shown to be significant with a sufficiently large sample size[2,3], there has recently been increasing interest in using difference tests to measure the size of sensory differences using Thurstonian theory[4]. Concurrent with this shift in perspective away from hypothesis testing and towards effect size estimation has been a desire to quantify the precision of these measurements[5]. Along with these developments, the importance of determining the size of consumer-relevant differences has also become apparent. For all these reasons, in this report we will consider the use of confidence intervals to help with the interpretation of sensory differences obtained through Thurstonian scaling and to make use of the precision of these estimates.

A value in converting choice outcomes to Thurstonian scaled estimates of δ, a standard measure of sensory difference, is that it allows a comparison of methodologies on a common basis and has provided empirically supported predictions about the relative power of different methods. This insight was recently used to support a switch from the Triangle test to the Tetrad test[6] and earlier provided a very satisfying explanation for the large difference in power between the 3-AFC and the Triangle test[7]. The experimental estimate of δ, called d', can be obtained easily from many methods and it is of interest to provide a way of calculating confidence intervals for δ and comparing the results to consumer-relevant specifications.

Scenario: You work for a national distiller and routinely conduct odor evaluations of rum, whiskey, vodka, and other liquors. You would like to obtain confidence intervals for δ values obtained from your traditional Triangle test data and compare them to those obtained from the Tetrad test. You are considering a switch in methodology if it proves beneficial to your sensory testing program. You have obtained 50 Triangle test judgments from experienced difference testing panelists for three rum variants compared to a gold standard reference. In order to compare methodologies, you also conducted three Tetrad tests with the same sample size – the results are shown in Table 1. The instruction in the Triangle test is to select the most different sample from three samples, in which two of them are putatively identical. The instruction in the Tetrad test is to evaluate two pairs of samples in which there are putatively identical samples within each pair and to group them into two groups of two based on similarity. Note that both methods have guessing probabilities of 1/3 and that your data fall in line with the general result that choice proportions from Tetrad testing are typically higher than those from Triangle testing[8]. Binomial tests on these data show that the null hypothesis that the true choice probability is 1/3 is rejected at the 95% level supporting the conclusion that all of the variants are different from the gold standard. However, the mere establishment of differences from the gold standard does not address the issue of the consumer relevance of the detected differences.

	Triangle Test	Tetrad Test
Variant 1	28	36
Variant 2	26	33
Variant 3	25	31

Table 1. Counts of correct responses in the Triangle and Tetrad tests for three rum variants with a sample size of 50 triads or tetrads.

Confidence Intervals for Binomial Data: One simple and almost universally used method for calculating confidence intervals is based on the normal approximation to the binomial. Confidence intervals computed this way are called Wald intervals. The Wald interval has been soundly criticized in the statistics literature. Brown *et al.* remarked, "Virtually all of the conventional wisdom and popular prescriptions are misplaced. The Wald interval is sufficiently poor in this problem that it should not be trusted unless npq is quite large."[9] Some alternatives have been suggested. These include likelihood-based intervals[10] and the Agresti-Coull interval,[9] which is easy to calculate and is more reliable than the Wald interval. Although the Agresti-Coull interval is sometimes conservative (i.e. longer than it needs to be), its simplicity recommends it and it will be used in this report.

To calculate a 95% Agresti-Coull confidence interval, it is convenient to approximate the two-tailed 95% bounds for the standard normal distribution as 2 rather than 1.96. Then if k is the choice count and n is the sample size, the Agresti-Coull confidence interval is:

$$\frac{k+2}{n+4} \pm 2\sqrt{\frac{(k+2)(n-k+2)}{(n+4)^3}}$$

If $k = 25$ and $n = 50$, the interval is (0.364, 0.636).

Confidence intervals for δ: Once a confidence interval for a choice probability from a sensory difference test has been worked out, it can be converted to a confidence interval for δ using the psychometric function for that difference test. These conversions can be made using tables[1] or the *IFPrograms*™ software. For instance, the confidence interval (0.364, 0.636) for the Triangle test corresponds to a confidence interval for δ of (0.59, 2.16). Note that the relationship between confidence intervals on choice probabilities and confidence intervals on δ will depend on the method. For example, Figures 1 and 2 show the relationships between δ and the probability of a correct response for the Triangle and Tetrad tests, respectively. These relationships are called psychometric functions. It can be seen that as δ changes within the region shown, the Tetrad psychometric function rises much more rapidly than the Triangle psychometric function. Thus, for a fixed interval in the choice probability, the corresponding interval in δ values is smaller for the Tetrad test than the Triangle test – this fact intuitively explains the greater precision of the Tetrad test. Figure 3 confirms this intuition

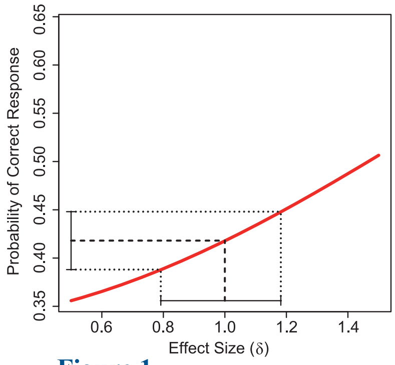

Figure 1.
Triangle psychometric function.

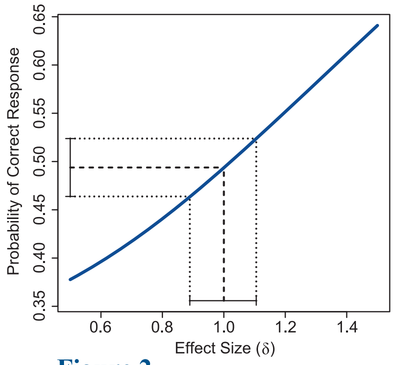

Figure 2.
Tetrad psychometric function.

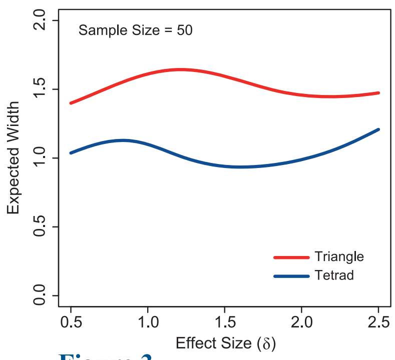

Figure 3.
Widths of confidence intervals.

by showing the average widths of the confidence intervals on δ generated by both methods for the same sample size.

Confidence Intervals for the Rum Data: Table 2 shows the Agresti-Coull confidence intervals for the three comparisons in Table 1 using the Triangle and Tetrad tests. Table 3 shows the corresponding confidence intervals on δ, and from this last table it can be seen that the confidence intervals for the Tetrad test are shorter than those for the Triangle test, confirming the greater precision of this method.

	Triangle Test	Tetrad Test
Variant 1	(0.420, 0.691)	(0.579, 0.830)
Variant 2	(0.383, 0.655)	(0.518, 0.778)
Variant 3	(0.364, 0.636)	(0.478, 0.744)

Table 2. Agresti-Coull confidence intervals for the data in Table 1.

	Triangle Test	Tetrad Test
Variant 1	(1.01, 2.45)	(1.29, 2.23)
Variant 2	(0.75, 2.26)	(1.09, 2.00)
Variant 3	(0.59, 2.16)	(0.94, 1.87)

Table 3. Agresti-Coull confidence intervals converted to *d'* values. Note that the Tetrad intervals are shorter than the Triangle intervals.

Consumer Relevance: Now that confidence intervals for your experiment have been constructed, it is useful to compare your results to a consumer-relevant action standard. Assuming that a δ of 1 has been established as consumer relevant, your current results can now be considered for consumer-relevance. Figure 4 plots the results from Table 3 against an action standard of 1. In this figure you see that, according to Tetrad testing, the sensory differences for variants 1 and 2 lie above the action standard, indicating that the differences would be noticeable to consumers. On the other hand, while your Triangle testing shows that the sensory difference associated with variant 1 lies above the action standard, Triangle testing does not provide similar confidence in the case of variant 2. It is worth noting that, although variant 3's confidence bounds include the action standard for both tests, its lower bound from Tetrad testing shows that it is quite close to the standard. Thus, if this variant involves a business-critical substitution, you may wish to conduct further testing to be sure that the change is not noticeable to consumers.

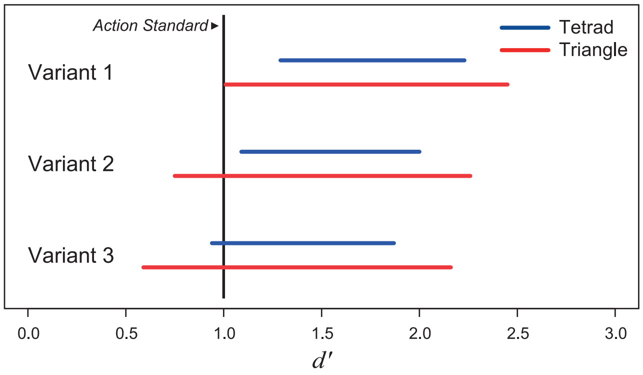

Figure 4. Confidence intervals for the Triangle test and the Tetrad test relative to an action standard of δ = 1.

Conclusion: Confidence intervals for binomial data can be converted to confidence intervals for δ, a standard measure of sensory difference. When such confidence intervals are conducted using the Triangle and Tetrad tests, the intervals are shorter for the Tetrad test than for the Triangle test. These shorter intervals justify a switch from Triangle to Tetrad testing based on the greater precision of the Tetrad test. In particular, we have seen that the Tetrad test provides greater confidence in diagnosing when sensory results exceed a consumer-relevant action standard.

References and Notes

1. *Tools and Applications of Sensory and Consumer Science* (pp. 130-159). Richmond, VA: The Institute for Perception.
2. Chew, V. (1977). Statistical hypothesis testing: An academic exercise in futility. In *Proceedings of Florida State Horticultural Society* (Vol. 90) (pp. 214–215). Lake Alfred, FL: FSHS.
3. Ennis, D. M. (1990). Relative power of difference testing methods in sensory evaluation. *Food Technology,* **44**, 114, 116, 117.
4. Ennis, J. M., Rousseau, B., and Ennis, D. M. (2014). Sensory difference tests as measurement instruments: A review of recent advances. *Journal of Sensory Studies,* **29**(2), 89–102.
5. Ennis, J. M. and Christensen, R. H. B. (2014). Precision of measurement in Tetrad testing. *Food Quality and Preference,* **32**(A), 98–106.
6. Ennis, J. M. (2013). The year of the Tetrad test. *Journal of Sensory Studies,* **28**(4), 257–258.
7. Frijters, J. (1979). The paradox of discriminatory nondiscriminators resolved. *Chemical Senses,* **4,** 355–358.
8. O'Mahony, M. (2013). The Tetrad test: Looking forward, looking back. *Journal of Sensory Studies,* **28**(4), 259–263.
9. Brown, L. D., Cai, T. T., and Dasgupta, A. (2002). Confidence intervals for a binomial proportion and asymptotic expansions. *The Annals of Statistics,* **30**(1), 160–201.
10. Christensen, R. H. B. and Brockhoff, P. B. (2009). Estimation and inference in the same-different test. *Food Quality and Preference,* **20**(7), 514–524.