SOME STATISTICAL CONSIDERATIONS FORS MALL SAMPLE EVALUATION IN TRIANGULAR TASTE TESTS

H. R. ROBERTS
C. H. McCALL, JR
R. E. THOMAS

本 論文은 現在 ASQC의 Electronic Div.의 News Letter 發行人이며. BOOZ ALLEN APRLIDE RESEARCH Inc의 副社長으로 있는 Chester H McCALL 博士가 宋泰昱 理事를 通해서 KSQC 會員들에게 討論의 機會를 마련하기 위해 보내준 것입니다. 이에 대한 意見이 있으시면 本學會로 보내 주시면 感謝 하겠읍니다.

A common problem with which taste panels are faced is that of distinguishing between two similar samples. The most widely used experimental designs for the solution of this problem are pair tests, duo-trio tests, triangle test, multiple comparison tests, and their various modification. All of the preceding tests have in common the fact that a decision is based on the proportion of correct judgments made by the panel members. In comparing the relative merits of the various experimental designs, apparently conflicting results have been obtained. Byer and Abrams (2) conclude that pair tests are superior to triangle tests. Hower, Harrison and Elder (5) call the triangle test "obviously more efficient" than the pair test. Gridgeman (4) concludes that "…pair tests and triangle tests are normally about equally powerful and appreciably superior to duo-triotests. "Hopkins and Giridge man (6) have shown that, in consideration of their respective powers, "triangular tests have a statistical advantage over duo-trios and pairs." In certain situations, experimental evidence would seem to indicate a preference preference for multiple comparisons (11, 12). The purpose of this paper is to consider a modification in the analysis of the triangular test, rather than to indicate the best experimental design (even if such were possible).

Although other than statistical considerations are important in most experimental situations, the statistical advantage of triangular taste teste tests is of sufficient merit to warrant further investigation of this technique. With this in mind, the authors have developed a statistical procedure for the evaluation of triangle test results which improves the analysys of these results, using as a criterion of improvement the powers of the conventional and modified test procedures.

Although the emphasis is primarily on the statistical aspects of this technique, it is recognized that the usual prodlems of coding bias, panel selection, fatigue, and other personal factors exist and must be dealt with by the experimenter. For discussions of these problems, the reader may consult the cita tions (1, 7, 9, 10, 13) contained in the reference list.

SOME THEORETICAI, ASPECTS

Let us assume, for the purpose of this paper, that the primary concern rests in the question: "Is there a detectable difference between samples A and B?" The intensity (3) of such a difference has been omitted from consideration, not only to simplify the analysis of the test but also due to the fact that practical situations exist for which the intensity is not a factor.

For example, a certain commodity, X contains 1% of an additive Y. It is desired to increase the percent of Y witnout substantially changing the flavor of X. If the falvor difference is detectable, the percent of Y will not be changed. However, if increasing the percent of Y to say 3% does not incur a detectable difference, the higher percent of Y may be used. In cases such as this, The intensity of The difference is of no consequence. The question is: Does a difference exist (regardless of degree)?

In a triangle taste test, wherein the panel attempts to determine which of 3 aliquotsis the"odd' or different aliquot, one would expect to pick the "odd" aliquot once in every 3 times on the average by merely guessing. In the notation of Gridgeman (4), let us define the probability of correct discrimination as P_d and the probability of picking the "odd" aliquot by guessing as P_G equals 1/3. Then the proportion of correct judgments in the long run is, by the law of compound probability:

$$P = P_d + P_G(1 - P_d) \text{ or } P = (1 + 2P_d)/3$$

Obviously, if there is no detectable difference in uhe samples, P_d is O and P is 1/3. However, if there is a readily detectable difference, P_d will be greater than O and hence P will be greater than1/3. This suggests the null hypothesis that the probability of correct discrimination is zero, which is the common uull hypothesis for sensory diffence tests. An alternative hypothesis that P_d is greater than zero is then also suggested by the fore going. Of course, inpractice P_d is seldom if ever specifiable *a priori*. This poses no difficulty in the present situation, however, if we consider the alternative hypothesis specifying a set of values of P_d (any P_d greater than zero) rather than some single value. Clearly, the null hypothesis implies no detectable difference whereas the alterna tive hypothesis implies the opposite. Also, for the sake of simplicity, we shall assume equalety the P_d for all judged of a Teavor for test is assumption, as well as for estimating P_d the reader is referred to Hopkins and Gridge man (6). (The authors are presently considering more general alternative hypothists wherein equality of P_d for all judges is not assumed.). In terms of a statistical test procedure, the null and alternative hypotheses are specified by

and $$H_0: P_d = O \text{ or } H_0: P = P_G = 1/3$$

$$H_1: P_d > O \text{ or } H_1: P > P_G = 1/3$$

respectively. The alternative hypothesis is then a composite hypothesis, i.e. it is consistent for more than one value of P. For testing the above hypotheses, a one-tailed test should be used. In other words, the critical region (that set of results which leads to the rejection of the null hypothesis and implies the acceptance of the alternative hypothesis) consists of only large values of P(P>1/3).

In any statistical test procedure, two main types of errors are to be considered. These two errors aretermed thetype I and type I errors, where the type I error is the error made in rejecting H. when in fact it is true and the type IIerror is the error made in failing to reject the null hypothesis when in fact it is not true. The above errors are commonly referred to as the producer's risk and the consumer's risk, respectively, since an error of the firsttype usually results in rejecting an acceptable item and an error of the second type usually results in the consumer being exposed to a nonacceptable item. In taste testing, committing a type I error results essentilly in a mistake by which the consumer is not adversely affected. However, committing a type II error may result in adverse consumer reaction. Thus, it is desirable to keep the probability of both errors small, but especially the probability of a type II errer. Unfortunately, the type II error. is frequently neglected in analyzing taste test results (14).

The probability of committing a type I error is denoted by α and the probability of committing a type II error is denoted by β. For a given test procedure, the critical region is determined by specifying α The optimum test situation is one in which, for a specified α, β is a minimum. However. for a fixed sample size, α and β vary inversely. Therefore, choosing $\alpha = O$ would make β a maximum and conversely. In experimental sisuations, however, an $\alpha = O$ is an impractical restriction; consequently, an α of .05 or .01 is conventionally used (8). Minimizing β is equivalent to maximizing $1-\beta$, which is called the power of the test. In a test situation where the alternative hypothesis is consistent for more than one value of P, the power of the test varies with the particular value of P which is selected. In this case, we obtain the power function which gives the power of the test for the various values of P between O and 1. Of course, the values of P which are of interest in the triangle test are $P = 1/3 (H_0)$ and $P > 1/3 (H_1)$. Values of P less

than 1/3 are not meaningful in this situation. The power function then appears to be one reasonable criterion for choosing between two test procedures.

Conventionally, the procedure for analysis of triangle test results uses the binomial expression:

$$(1.1) \qquad (P+Q)^n = \sum_{x=0}^{n} \binom{n}{x} P^r Q^{n-x}$$

where x is the number of correct determinations out of a total of n determinations, P (the probability of a correct determination) is specified by the null hypothesis and $Q=1-P$. For a specified α, the critical region is determined by solving the expression:

$$(1.2) \qquad \sum_{x=x_0}^{n} \binom{n}{x} P^x Q^{n-x} \leq \alpha$$

for x_0. In a given experiment, the null hypothesis is rejected whenever the number of successes is equal to or greater than x_0. Since the binomial is a discrete distribution (x assumes only integral values), (1.2) is expressed as an inequality rather than an equality. Unfertunately, unless n is sufficientry large, the term on the left side of (1.2) will vary from α by a great deal (see section on procedure). Although it may not be obvious without a numerical example, this in turn increases β and hence becreases the power of the test.

As a partial selution to this problem, the authors propose the use of the multinomial rather than the binomial distribution. Let us assume, for the sake of simplicity, that the experiment involves 3 replications or trials by each judge. (It should be realized that small-sample designs involving other groupings of judges and replications are notonly possible but of considerable interest. power considerations would be of importance in these cases also, as a means of choosing appropriate designs. However, the authors feel that consideration of these cases would tend to make the present paper too lengthy for the purpose intended.) There are then 4 posible results. A judge may make 3, 2, 1, or 0 corret decisions. Let:

x_1 be the number of judges making 3 correct decisions

x_2 be the number of judges making 2 correct decisions

x_3 be the number of judges making 1 correct decision

x_4 be the number of judges making 0 correct decisions

and let:

Π_1 be the probability of a judge making 3 correct decisions

Π_2 be the probability of a judge making 2 correct decisions

Π_3 be the probability of a judge making 1 correct decision

Π_4 be the probability of a judge making 0 correct decisions.

The Π_1 values are then determined by the following considerations: The probability of a judge making 3, 2, 1, 0 correct decisons out of 3 possible correct decisions, when the probability of a correct decision under the null hypothesis is 1/3, is given by the binomial expressions:

$$\Pi_1 = \text{probability } (r, n) = \binom{n}{r} P^r Q^{n-r}$$

where $r=3, 2, 1, 0$; $n=3$; $P=1/3$; and $Q=1-P=2/3$.

We then have:

$$\Pi_1 = \text{probability } (3, 3) = \binom{3}{3} (1/3)^3 \ (2/3)^3 \ (2/3)^0 = 1/27$$

$$\Pi_2 = \text{probability } (2, 3) = \binom{3}{2} (1/3)^2 = 6/27$$

$$\Pi_3 = \text{probability } (1, 3) = \binom{3}{1} (1/3)^1 (2/3)^2 = 12/27$$

$$\Pi_4 = \text{probability } (0,3) = \binom{3}{0} (1/3)^0 (2/3)^3 = 8/27$$

The null hypothesis could then be written as:

$$H_0: \Pi_1 = 1/27,\ \Pi_2 = 6/27,\ \Pi_3 = 12/27, \Pi_4 = 8/27$$

but this is exactly equivalent to writing:

$$H_0: P = 1/3.$$

Hence the simpler form will be used along with $H_1: P > 1/3$. The multinomial expression corresponding to (1.1) is then:

$$(1.3) \quad (\Pi_1 + \Pi_2 + \Pi_3 + \Pi_4)^K = \sum_{x_1} \frac{K!}{X_1! X_2! X_3! X_4!} \Pi^{x_1}{}_1 \Pi^{x_2}{}_2 \Pi^{x_3}{}_3 \Pi^{x_4}{}_4$$

where k is the number of judges on the panel and $\Sigma x_1 = k$. For a specified value of α, (1.3) may be summed over the values of x_1 which tend to discredit the null hypotesis and such that the sum is less than or epual to α. The relatively greater power of this method will be illustrated in the section onprocedure.

PROCEDURE

For illustration, let us use 4 judges and 3 replications. There are then 12 possibilities of a correct determination. If no readily detectable difference exists between the samples, approximately 4 correct decisions (1/3 of 12) would be expected. The reader should realize, however, that the proportion of correct judgments will be close to 1/3 only if the experiment is repeated a large number of times. In single experiments, the results will vary about 1/3 by a chance amount when there is no readily detectable difference. The hypotheses are:

$$H_0: P = 1/3 \text{ and } H_1: P > 1/3.$$

Let us select an α of. 05. This means that a certain set of resultt will be assumed to indicate a real difference between samples whenever the chance of such results is less than 5 in 100 under the nullhypothesis. Under these conditions we shall, of course, expect to be wrong 5 times out of 100 in the long run. The two forms of analysis follow:

Binomial:

Here we have:

$$(1.4) \qquad \sum_{x=0}^{n} \binom{n}{x} P^x Q^{n-x} = \sum_{x=0}^{12} \binom{12}{x} (1/3)^x (2/3)^{12-x}$$

Then (1.2) becomes:

$$(1.5) \qquad \sum_{x=x_0}^{12} \binom{12}{x} (1/3)^x (2/3)^{12-x} \leqq 05$$

which is found to be consistent for $x_0 = 8$ by consulting the binomial tables(15). Therefore in this test, the null hypothesis would be rejected whenever 8 or more correct determinations were made. It should be pointed out, however, that the left side of (1.5) isactually equal to. 019 for $x_0 = 8$.

Multinomial:

The terms of the multinomial are given by:

$$(1.6) \qquad \sum_{x1} \frac{4}{X_1! X_2! X_3! X_4!} (1/27)^{x1} (6/27)^{x2} (6/27)^{x3} (8/27)^{x4}$$

These are arranged by ordering the number of succeees (0—12)and the tererms correspondin; to each munber of correct determinations. (Actually then, the first stage ordering of these terms is abinomial ordering and the multinomial ordering is the second stage. Howeve-, sincethe improvement results from the multinomial brea'k-down and ordering, the whole procedure ittermed multinomial break down and ordering, the whole procedure is termed multinomial.) The reader will notice that for each number of corrct determinations shere may be several multino-

mial terms, These terms, each corresponding to the same number of correct determinations, are then ordered by considering the probability of their occurrence under H_0 and H_1, i.e., the terms least likely under H_0 are placed nearest the rejection region In ordering the multinomial terms several possibilities exist, but the method used herein is felt to be the most practical since the probabilities of term sin the rejection region consistently increase with increasing values of P. This illustrates the practical utility of the multinomial in contrast to the bnomial. The binomil number of considers only the Corret determinatins, Wherease the multinemial also considers the ways in which a given number of correct determinations could originate.

For the case considered here, the inequality:

$$(1.7) \qquad \sum_{x_1} \frac{4!}{X_1!\,X_2!\,X_3!\,X_4!}\,(1/27)^{x1}(6/27)^{x2}(12/27)^{x3}(8/27)^{x4} \leqq 05$$

is satisfied for 7 correct decisions or more providing that the 7 correct decisions are a result of:

$$(1.8) \qquad x_1=2, x_1=0, X_3=1X_4=1$$
$$x_1=1, x_2=2, x_3=0, x_4=1$$
and
$$X_1=0, X_2=3, X_3=1, X_4=0$$
or
$$X_1=1, X_2=1, X_3=2, X_4=0$$

A problem, uniquc (among the results included) to 12 possible correct decisions, is encountered in this case. Both of the last two results have the same probability of occurrence; but, if both are included in the rejection region (1.7) becomes approximately .066 and the inequality is not satisfied. Two possible solutions are to include only one of the results and to include neither of the results. Including only one, (1.7) is approximately .047. Including neither, (1.7) is approximately .027. However, in both cases the resultant power is greater than under the binomial. In the binomial case, the corresponding expression is approximately .019. If one of these two results were obtained experimentally, an objective decision could be made based on the combination of α and β which was most advantageous to the experimenter. For example, the maximum α and minmum β would be obtained if both results were included and the converse would be true if neither were included. For the purposes of constructing the multinomial power function, one of the two results was included in the rejection region.

Tables 1 and 2 give the results which lead to a rejection of the null hypothesis for an α of .05 and .01, respectively, when 1, 2, 3, 4, 5, or 6 judges and 3 replications are used. For example, the results derived above are listed in Table 1 under 4 juges. There we find that 8 to 12 correct decisions and additicnal results listed in (1.8) lead to a rejection of the null hypothesis for an α of .05.

TABLE 1

.05 Critical regions for multinomial triangular taste tests1 through 6 judges:

3 replicates per judge

Number of judges:	1	2	3	4	5	6
Critical regiong:	3 correct	5-6 correct	6-9 correct	8-12 correct and	9-15 correct and	10-18 correct and
				X_1 X_2 X_3 X_4	X_1 X_2 X_3 X_4	X_1 X_2 X_3 X_4
				2 0 1 1	2 1 0 2	3 0 0 3
				1 2 0 1	2 0 2 1	1 3 0 2

and 0 4 0 1 2 1 1 2

0 3 1 0 2 0 3 1

or

1 1 2 0

TABLE 2

01 Critical regions for multinomial triangular taste tests—1 through 6 judges; 3 replications per judge

Number of judges:	1	2	3	4	5	6
Critical regions:	•	6 correct	7-9 correct and	9-12 correct and	10-15 correct and	12-18 correct and
			X_1 X_2 X_3 X_4	X_1 X_2 X_3 X_4	X_1 X_2 X_3 X_4	X_1 X_2 X_3 X_4
			2 0 0 1	2 1 0 1	3 0 0 2	3 1 0 2
				2 0 2 0	2 0 3 0	3 0 2 1
				0 4 0 0		1 4 0 1
						0 5 1 0
						2 1 3 0
						2 2 1 1

• No rejection possible at. 01.

Power Functions-Four Judges

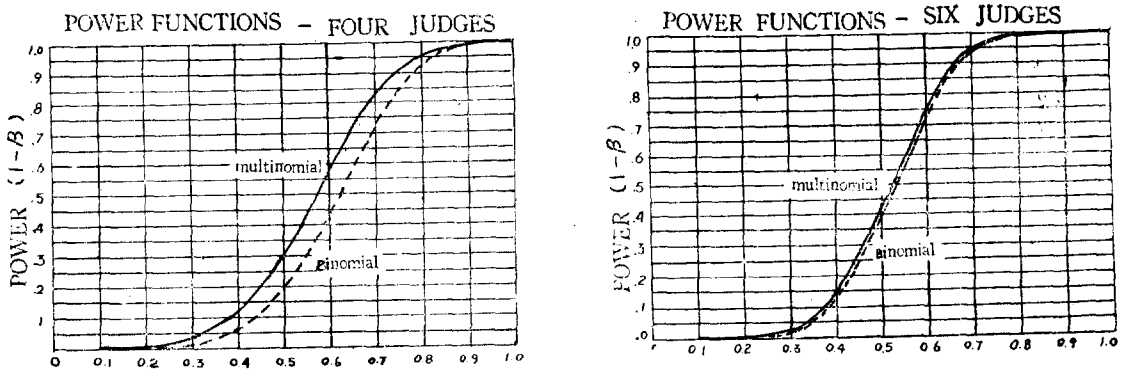

Fig. 1.

POWER OF THE

The power curves for the binomial and multinomial are given in Figure 1, for an α of .05 and for 4 and 6 judges. There, the power of the tests $(1-\beta)$ is plotted against the various values of P. Since β is the probability of not rejecting H_0, when in fact some value of P consistant with H_i is true, $1-\beta$ may be arrived at by summing, for the various values of P, the probabilities of the results which fall in thecritical region (See Tables 1 and 2). This has been done for values of P less than 1/3 also, in order to complete the power function. As pointed out previously, however, these values of P have on interpretationt for the test in question. As evidenced

by Figure 1, the power of the multionmial is consistently greater than that of the binomial for the two cases cited. (Similar results follow for an α of .01). For more than 6 judges and 3 replications, the two power curves practically coincide. For these situations then, there is little advantage in using the multinomial rather than the binomial from the point of view oof power, there is a sufficient number of small sample cases to warrant the application of the results derived from consideration of the multinomial.

LITERATURE CITED

1. BENNETT. G., SPAHR, B., and Dodds, M. The value of training a sensory test panel. *Food Technol.*, 10, 205 (1956).
2. BYER A. J., AND ABRAMS, D. A comparison of the triangular and two-sample tastetest methods. *Food Technol.*, 7, 185 (1953).
3. DAVIS, J., and HANSON, H sennsory test methods. I. The triangle intensity (T-I) and related test systems for sensory analysis. *Food Technol* 8, 335 (195. 4)
4. GRIDGEMAN, N. Tasts comparisons: two samples or three? *Food Technol.*, 9, 148(1955)
5. HARRISON, S., AND ELDER, L. W. Some applications of laboratory taste testing. *FoodTechnol.*, 4, 434 (1950).
6. HOPKINS 'J W., and GRIGEMAN, N. T. Compaative sensitivity of pair and triad flavor intensity' difference tests. *Biometrics*, 11, 63 (1955).
7. ISHLER, N., LAUE, A., and JANISCH, A. Reliability of taste testing and consumer testing. methods. II Code bias in consumer tepts *Food Techol.*, 8, 389 (1954).
8. KRAmeR, A., AND DITMAN, L. A simplified variables taste panel method for detecting flavor dhanges in vegetables treated with pesticides. *Food Techno'.*, 10, 155 (1956).
9. LAUE, E., ISHLER, N., AND BULLMAN, G. Reliability of taste testing and consumer testing methods. I. Faligue in taste testing. *Food Technol.*, 8, 389 (1954)
10. MACKEY, A., and JONES' P. Selection of members of a food tasting panel: Discernment of primary tast es in waiter *Food Techno!.*, 8, 527 (1954).
11. MAHONEY' C., STIER, H., and Crosby, E. Evaluation of flavordifference in canned foods due to the application of pesticides to the processing crop. I. Studies leading to the development ofa simplified procedure for making flovor difference tests. *Food Techno'.*, 11 (9) Insert, p. 29 (1957).
12. MAHONEY, C. ' STIER, E. Evaluation of flaver differences in canned foods due to the application of pesticides to the processing crop. II. A suggested simplified procedure for making flaver difference evaluations. *Food Technol.*, 11, (9) Insert, p. 37 (1957).
13. MITCHELL, J. Time-errors in the paired comparison taste preference test. *Food Technol.*, 10, 218 (1956).
14. RADKINS, A. Some statistical considerations in organoleptic research: triangle, paired, duo-trio tests. *Food Research*, 22, 259 (1957).
15. UNITED STATES DEPATMENT OF COMMERCE. *Tables of the Binomial Probability Distribution.* 1950.