

---

# Statistics for food science III: sensory evaluation data. Part B – discrimination tests

---

*John A. Bower*

---

## The author

John A. Bower is a Lecturer in Food Science in the Department of Applied Consumer Studies, Queen Margaret College, Edinburgh, UK.

---

## Abstract

Describes statistical methods applied to sensory discrimination tests. Illustrates binomial and chi-square statistical analysis and discusses similarity testing, power and replication in discrimination testing.

In part A of this series on statistical analysis of sensory data, published in the November/December 1995 issue of this journal the nature of significance testing and the statistical characteristics of such data were detailed. It is now time to examine some common sensory methods and data forms with the purpose of determining appropriate statistical methods.

## Which factors influence selection of statistical tests for different sensory evaluation methods?

Statistical tests are based on certain assumptions regarding data form and parent population characteristics, etc. If these assumptions are false the conclusions reached may be weakened and the test rendered invalid. Several factors must be taken into account. The level of measurement and the nature of the population distribution from which the data come are major points which often decide whether a *parametric* or *non-parametric* method is employed. The relationship of the individual treatments examined decides whether a dependent or independent sample test is appropriate. This can be a difficult issue in some experimental designs but, in sensory evaluation, dependent treatments may occur when all samples are examined by the same panel members. Independent treatments may occur when comparing data from similar samples generated by different panels, as might be done in storage trials, or a product survey sited in different locations. It is possible to perform a *one sample test* where sample results are compared with a known population parameter. If the number of treatments compared is limited to two then a *two sample test* (independent samples) or a *paired test* (related samples) is appropriate. For three or more treatments, methods based on *analysis of variance* are used. Some sensory techniques, such as discrimination testing and ranking, are limited in the choice of test by the nature of the data they generate. It is also possible to analyse by more than one type of test or reduce data in form to analyse by a simpler or less powerful test. Also important is the objective of the experiment in respect of the statistical inferences which are to be made – are these to apply to the food, or to the panel, or to both? If the sensory panel is used as an instrument to measure some aspect of quality in representative samples of food products,

the inferences will relate to the larger population of such products. In other circumstances the intention may be to provide information concerning a consumer population via a random consumer sample of panellists. These points must be considered before selecting a statistical method.

### Which statistical analysis is appropriate for data from sensory discrimination tests?

Data generated by sensory discrimination tests are usually categorical or nominal in nature. Results are discrete, taking only integer values. No assumptions are made regarding the nature of the population from which the data come and non-parametric methods of analysis are commonly used. Possible tests include the binomial test and the chi-square test. Owing to their non-parametric nature, these tests can tell us whether or not samples differ significantly, but cannot provide information regarding the extent of difference. The *binomial test* is almost universally used, and binomial probabilities are used to generate the well-known tables of "correct" or "agreeing judgements" which can be consulted to determine significance.

### Application of the binomial method to sensory discrimination test data

Imagine a product development situation where the required amount of a fruit flavour additive in a fruit-based dessert product has to be determined. The initial objective is to find a level which shows a detectable difference. Laying aside arguments as to whether or not a difference test should be employed to answer this question, assume that a paired comparison test is used. The supplier has specified a minimum level of use for the flavour and this quantity is used in the experiment. Sets of two samples from formulations of the food product, one with the new flavour (A), the other without (B), are presented to ten taste panel members. They are asked to select the sample which they deem to have the stronger fruit flavour, using a forced choice paired comparison difference test. Eight choose A and two choose B. On the surface, eight out of ten would appear to be convincing evidence that the flavour was definitely detectable – but is it *statistically* significant or could such a result happen by chance? If there

was no difference in flavour what would the result be? The binomial test can determine the probability of the result happening by chance alone – the lower the chance probability the higher the probability of a real difference in flavour given such a result. The statistical procedure follows below.

### Statistical procedure for significance testing

A concise approach is recommended as it clarifies the purpose and conclusion of the statistical analysis. First, the objective of the experiment and the result obtained are clearly stated:

*Objective:* To establish overall difference in flavour

*Result:* Total panel numbers = ten  
Panellists selecting formulation A = eight  
Panellists selecting formulation B = two.

The statistical procedure can now commence with a statement of the hypotheses:

- *Null hypothesis ( $H_0$ ):* There is a 50:50 proportion of responses.
- *Alternative hypothesis ( $H_1$ ):* There is a proportion of responses in favour of A.

Note the manner in which the hypotheses are stated – this has a bearing on later conclusions. The significance level and the scope of the test must now be stated:

Significance level ( $\alpha$ ): 5 per cent ( $P = 0.05$ )  
one-tailed test.

The most conservative 5 per cent level is selected, representing a probability ( $P$ ) of 0.05 of rejecting a true null hypothesis (probabilities can be expressed on a scale of 0 to 1 or 0 to 100 per cent). In order to be more certain that a true null hypothesis is not rejected then a significance level of 1 per cent could be selected, i.e. the type I error risk level ( $\alpha$ ) could be lessened. The alternative hypothesis is directional and it is possible to predict the outcome – if the null hypothesis is false there is reason to expect that the sample with added flavour will dominate the other in fruit flavour strength – hence a *one-tailed* test is used. This assumes that the experimenter is *absolutely sure* that the only alternative to no difference in flavour is that flavour A is the stronger. If this expectation is not tenable, or the experimenter wishes to be conservative, then a *two-tailed test* is more appropriate and the alternate hypothesis would consider one

of two possibilities, namely A greater than B or B greater than A.

The statistical test to be employed and the circumstances are now specified:

Statistical method: the binomial test – matched samples; independent assessments.

The conditions of the experiment are also important for the assumptions of the statistical analysis – a pair of samples which are matched (in terms of preparation procedure, etc.) and assessed by the same group of tasters, tasting each sample once, all scores being independent, i.e. the probability of each selection should not be influenced by the others. It is also assumed that correct experimental control has been implemented for all aspects of the test. The binomial statistic is obtained by use of the general term or by summation of specific terms from the binomial expansion :

$$\text{Binomial expansion} = (p + q)^n$$

where:

- $p$  = chance probability of each trial succeeding;
- $q$  = chance probability of the trial failing and  $p = (1 - q)$ ;
- $n$  = number of trials (independent selections).

For example, if  $n = 3$  the expansion becomes  $p^3 + 3p^2q + 3pq^2 + q^3$ .

For large numbers of trials (e.g. single selections by more than 50 panellists) a normal approximation to the binomial distribution can be used in place of the expansion or the binomial term. Binomial statistics deal with event probabilities, i.e. the chances of success or failure for events or trials. In the circumstances of sensory discrimination, the “trial” referred to is the selection procedure which each panellist performs; thus there are ten trials in the example above. Analysis of the decision process through which individual panellists go can aid understanding of the statistical test and gives an additional insight into the sensory methodology.

On presentation of the samples the panellists must decide which of the two has the stronger flavour. If there is an obvious difference a decision will be reached without difficulty. For smaller differences panellists are forced to concentrate and choose one sample over the other. If they cannot detect a difference the test does not permit a “no difference” choice. Note that giving such options changes

the whole framework of the statistical procedures and the nature of the discrimination test[1]. For panellists who cannot detect a difference the test reduces to a guess or “pick one at random”. In theory each sample should have an equal chance of being selected – rather like mentally flipping a coin to decide. In the context of sensory evaluation this may or may not be the case, as presentation and other factors could interfere and negate the equal chance choice. Assuming adequate control over these factors, the chance probability of any one sample being picked is 50 per cent or 0.5, and the chance of not being picked is also 50 per cent as there are two samples. These represent the chances of success or failure referred to above. We are in effect assuming that the null hypothesis is true for these circumstances.

The binomial expansion will give the chance probability for all possible outcomes e.g. if the difference is not detectable by the panel the chance of ten panellists choosing A is given by the first term:

$$(0.5)^{10} = 0.00098 \text{ or } 0.098 \text{ per cent.}$$

The expansion can be extrapolated out to give the chance probabilities for each of the ten possible outcomes to the sensory test (Table I).

The “50:50 split” result is the most probable event and the other outcomes are less likely and are symmetrical about this point. Low probability values mean that it is extremely unlikely that such an event would occur by chance. Consequently, if such a result were obtained it is highly likely to be

Table I All possible outcomes and binomial probabilities of a paired comparison difference test for ten panellists

Outcome of experiment (event)		
No. of panellists choosing A	No. of panellists choosing B	Chance probability (binomial) <sup>a</sup>
10	0	0.00098
9	1	0.0098
8	2	0.044
7	3	0.117
6	4	0.205
5	5	0.246
4	6	0.205
3	7	0.117
2	8	0.044
1	9	0.0098
0	10	0.00098

Note: <sup>a</sup>Values are rounded off for simplification

because of the presence of a real difference. It is not recommended to attempt to calculate out the full expansion manually – appropriate software can achieve this, but the process is made much easier for the sensory analyst as probability and event tables are available in the literature [1,2]. The terminology used in these tables refers to combined event probabilities, i.e. in the example above we require not just the probability of “eight out of ten”, but rather that of “eight or more out of ten”. The probabilities for eight, nine and ten out of ten must be summed. Thus all the probabilities are not required – only those at and above the sample choice which dominates (or either one if panel choices are equal). Thus in order to establish the probability of getting eight or more out of ten in favour of A, addition of terms eight, nine and ten is required:

$$P = 0.044 + 0.0098 + 0.00098 = 0.0548.$$

Consulting probability tables [1] for the circumstances gives a chance probability of 0.055 which agrees closely with the calculated result. This constitutes the *test statistic* which can now be compared with the significance level:

$$\text{Test statistic: } P = 0.055$$

$$\text{Significance level: } P = 0.05.$$

There is no critical value as such, but for the result to be significant the test probability would have to be equal to or less than the assigned significance level. Based on this comparison the final conclusion can be made:

The probability of the result occurring as a result of chance alone was greater than  $P = 0.05$ . Hence the null hypothesis is retained.

Thus the data gathered by the experiment have not provided sufficient evidence to reject the null hypothesis and establish that there is a difference in strength of flavour. This concurs with the conclusion reached on consultation of paired comparison test event tables [2] which reveal that eight out of ten would be insignificant at the 5 per cent level (at least nine out of ten would be required for significance). These latter tables are easy to use but some formats may give an all or nothing impression to the uninitiated user, i.e. nine out of ten agreeing judgements is significant at the 5 per cent level and anything less is deemed non-significant. The probability value, although non-significant, is close to 5 per cent and may provide useful information

for subsequent work. If the  $\alpha$  level were set at 10 per cent, or even 6 per cent, the result would be significant – this underlines the point regarding slavish adherence to the 5 per cent level.

The conclusion does not provide information as to why the difference was not statistically significant. It could be because of one or more factors: a true null difference, lack of sufficient magnitude of difference, lack of panellist discrimination acuity or poor experimental control. The statistical procedure is summarized below:

- $H_0$ : Proportion of responses is 50:50.
- $H_1$ : Responses (A) > responses (B).
- $\alpha$ : 5 per cent;  $P = 0.05$ .
- One-tail/two-tail: one-tail.
- Method: binomial test.
- Critical region:  $P \leq 0.05$ .
- Calculated statistic:  $P = 0.055$ .
- Conclusion:  $H_0$  is retained.

### Binomial statistics and other discrimination tests

The example above was limited to one discrimination test, but other tests are dealt with in a similar manner. The chance probabilities are the same for the duo-trio test (50:50) whereas for the triangle test where a choice from three samples occurs the chance probabilities become 33⅓ per cent (success) and 66⅔ per cent (failure). Increasing the number of stimuli for selection will lessen the chance odds – the two out of five test reduces the success probability to 10 per cent. Fewer correct scores are required for significance in these latter tests because of the lower chance probabilities; for example, in the triangle test eight out of ten is significant with  $P = 0.01$ . This apparent advantage is offset by the possibility of sensory confusion caused by the increased number of “tastings” and the complexity of the task. The decision process required for the triangle test differs from that of the paired comparison and the probability advantage should not be confused with the relative sensitivity of the tests. A more sensitive discrimination test will be able to detect given differences with greater probability and has greater statistical power. The basic form of the triangle test has been shown to be lacking in this respect [3]. The triangle test and duo-trio are one-tailed in nature and are decided by the number of correct responses received. A paired preference test is two-tailed

– it is not possible to predict which direction preference will take if there is a detectable difference. The binomial probabilities would need to be doubled for two-tailed tests.

### Does an insignificant difference test result tell us that the products are the same?

In the sensory experiment above the objective was to establish the presence of a detectable difference. Use of discrimination tests in the quality assurance situation may have a different objective – to establish that there is no difference between product samples. The conclusion of the fruit flavour experiment was that there was a lack of a significant difference. This can be re-phrased as lack of a detectable difference and cannot be equated with stating that there is no difference. To provide this information a variation of difference testing which has been called *similarity testing*[4] can be used. It is necessary to specify the type II error level or  $\beta$  risk, i.e. the risk of declaring two products indistinguishable when they are in fact different. Second, an acceptable estimate of a proportion of the population who would detect a difference must be specified. This enables a conclusion of a different form to be reached. For example, with  $\beta$  set to 1 per cent and a proportion of discriminators estimated at 25 per cent, a similarity test is run using a triangle test. On receipt of sufficiently low correct responses, the experimenter would be able to conclude with 99 per cent confidence that no more than 25 per cent of the consumer population would detect a difference. Further details and tables for this type of power analysis approach are given in the literature[4-6].

### What if another type of statistical test is used?

The *chi-square test* can also be used for data of this type – it compares the actual or observed frequency of response with the expected response level, assuming no difference between the samples. Certain statistical considerations affect the use of this method – a minimum incidence of five (ideally ten) for each expected frequency, with a total number of observations above 40 (ideally 50) is recommended. Also when the number of categories (possible sample selections) is limited to two a correction factor may have to be

applied. Statistical texts differ in the exact applicability of these rules and caution is advised in unguided use of the chi-square method with discrimination test data.

### Chi-square method for discrimination tests

The chi-square test calculates the probability of getting the experimental result on the basis that the null hypothesis is true.

#### Chi-square formula (uncorrected)

$$\chi^2 = \sum \left[ \frac{(O - E)^2}{E} \right]$$

where

$O$  = observed frequency

$E$  = expected frequency

$\Sigma$  = “sum of”.

Once the correction factor is applied[7,8] the formula is adjusted and becomes simpler for a paired test.

#### Chi-square formula (corrected: two categories)

$$\chi^2 = (\text{Absolute difference between sample selections} - 1)^2 / \text{total number of selections.}$$

For example:

$$\chi^2 = (|8-2| - 1)^2 / 10 = 2.5.$$

The statistical procedure for the chi-square test on the fruit flavour example is summarized below:

- $H_0$ : Frequency is 50:50.
- $H_1$ : Frequency (A) > frequency (B).
- $\alpha$ : 5 per cent;  $P = 0.05$ .
- One-tail/two-tail: one-tail.
- Method:  $\chi^2$ (adjusted – Yates’ correction).
- Calculated  $\chi^2$ : 2.5.
- Tabulated  $\chi^2$ : 2.71 (1df,  $P = 0.05$ , one-tail).
- Conclusion:  $H_0$  is retained.

There are two categories in the test – the choice incidence for formulations A and B. If there is no difference in flavour, i.e. the null hypothesis is true, then the expected frequency would be 50:50 or five for each category – five selecting A and five selecting B. This did not occur and the observed frequency was eight for category A and two for category B.

Using the corrected formula (manually or by use of software) for these data gives a test statistic of 2.5. This must be compared with a tabulated critical value from  $\chi^2$  tables which

give the maximum values of  $\chi^2$  obtainable by chance if the null hypothesis is true. Applying the same circumstances as used for the binomial test, namely significance level (5 per cent), one-tailed test and the degrees of freedom (df) of the test (the number of categories minus one), consulting tables[8] gives a critical value of 2.71. The obtained chi-square value is not greater than the critical value. Hence the null hypothesis is retained – this is the same conclusion as reached by the binomial test.

Inappropriate use of chi-square could possibly produce invalid conclusions; for example, application of the non-adjusted formula would give a chi-square value of 3.6 which would be significant. Relatively, the uncorrected chi-square test is more powerful than the binomial test – i.e. on statistical analysis it has a higher probability of rejecting the null hypothesis. This means that for a given experiment the binomial test could require a larger difference between the pairs to obtain a significant result – at 50 trials chi-square would give significance for 31 or more selections of sample A over B; the binomial method would require at least 32. This distinction does not usually apply when the adjusted formula is used. Both methods can be used to aid decision making or to emphasize the  $\alpha$  or  $\beta$  risk[9]. Although the difference between the two tests is small, the risk ( $\beta$ ) of accepting a false null hypothesis is less with chi-square (uncorrected). If one test shows significance and the other does not, further testing may be required, i.e. do not accept that there is a real difference until the result is unanimous. The chi-square test has wider applicability and can be extended to more than two categories; it is used in analysis of data generated by the “A” - “not A” difference test[10].

### What effect has panel size on the statistical analysis?

In the above example there were ten panellists; a minimum of seven (expert) and 20 (selected) are recommended for paired comparison directional tests[2]. Larger panel sizes mean that a lower incidence of agreeing or correct judgements is required for significance. A panel of ten requires a minimum proportion of 90 per cent, whereas a panel of 100 requires 59 per cent ( $P = 0.05$ , one-sided) for a paired directional test. With large panel sizes organizational difficulties and the potential for

preparation and data transcribing errors may increase. Also, the number of responses for significance is only slightly above chance – 59 out of 100 (significant) is very close to the 50:50 split (chance). Thus with large sample (panel) sizes even marginal differences can be statistically significant. In the sensory evaluation context this statistical consideration could undermine the practical value of results. Possible reinforcements which could be applied are to increase the  $\alpha$  level or to use another discrimination test which has a lower chance probability. In the case of similarity testing, larger panel sizes are required to minimize the  $\beta$  risk. Panel size requirements can become unrealistically large if both types I and II error levels are minimized[6]. Ultimately the sensory analyst must balance up the statistical issues, the panel's nature and ability and the objective of the experiment.

### What if the discrimination test is replicated?

In the fruit flavour example each panellist performed the test once. Replicate tasting is analysed in a similar manner, but it may be incorrect procedure to combine panellists and replicates as separate results. Such a procedure may violate the required statistical independence. Thus, if ten panellists taste in triplicate, it does not mean that 30 independent trials have been made. Possible solutions to this problem include analysing the proportion of responses in the replicate sets to establish if they are significantly different[1] (e.g. by chi-square) – if insignificant, the responses can be pooled and the usual binomial tables used. Another approach is to analyse the replicates of each panellist first to give a single decision per individual, i.e. analyse the replicates independently. The independent decisions can then be analysed as above. The problem then is on what basis are the replicates to be analysed? Using the binomial method to decide whether or not an individual's proportion of correct or agreeing responses is significant, the number of replicates must be at least five (paired difference), i.e. even four out of four correct responses is not statistically significant ( $P = 0.05$ , one-tail). This level of replication and above is used in selection of panel members, but here adequate acuity may be judged by an arbitrary level of accuracy – above 50 per cent success rate, for example. Similar simple decision rules can be used as an

analysis process for small numbers of replicates, namely, if a panellist selects formulation A over formulation B three times out of four (75 per cent agreement), then count this result as “selecting A”. Replicate sessions, rather than tasting all at the one session, may provide the degree of independence. Replicate tasting, if done independently, can increase the confidence with which the panel organizer views the results.

The common sensory difference tests (paired, duo-trio, triangle) in the form above test statistical hypotheses regarding above-chance response proportions caused by the presence of detectable differences. It is possible to use them to measure the magnitude of the difference between samples by a procedure based on psychometrical functions of the sensory response stimuli [11]. Thus the proportion of responses in these non-parametric tests can derive a parametric measure of sensory discrimination. A non-parametric signal detection procedure (the R-index) has also been advocated for a similar purpose [12].

For all types of discrimination test the sensory analyst must beware of overstating the conclusions of the statistical analysis. The use of small panel numbers, panellists who are trained and who have a knowledge of the product, may mean that the results cannot be taken as representing the views of consumers.

## References

- 1 Stone, H. and Sidel, J.L., *Sensory Evaluation Practices*, 2nd ed., Academic Press, London, 1993, pp. 173-95.
- 2 British Standards Institution (BSI), *BS 5929: Methods for Sensory Analysis of Foods Part 2: Paired Comparison Test*, British Standards Institution, London, 1982.
- 3 Ennis, D.M., “Relative power of difference testing methods in sensory evaluation”, *Food Technology*, Vol. 44 No. 4, 1990, pp. 115-17.
- 4 Meilgaard, M., Civille, G.V. and Carr, B.T., *Sensory Evaluation Techniques*, 2nd ed., CRC Press, Boca Raton, FL, 1991, chapter 6.1, part IX.
- 5 Macrae, A.W., “Confidence intervals for the triangle test can give reassurance that products are similar”, *Food Quality and Preference*, Vol. 6, 1995, pp. 61-7.
- 6 Schlich, P., “Risk tables for discrimination tests”, *Food Quality and Preference*, Vol. 4, 1993, pp. 141-51.
- 7 Gacula, M.C. and Singh, J., *Statistical Methods in Food and Consumer Research*, Academic Press, Orlando, FL, 1984, pp. 334-6.
- 8 Amerine, M.A., Pangborn, R.M. and Roessler, E.B., *Principles of Sensory Evaluation of Food*, Academic Press, New York, NY, 1965, pp. 440-5.
- 9 O’Mahony, M., *Sensory Evaluation of Food Statistical Methods and Procedures*, Marcel Dekker, New York, NY, 1986, pp. 94-5.
- 10 British Standards Institution (BSI), *BS 5929: Methods for Sensory Analysis of Foods Part 5. “A”-“not A” Test*, British Standards Institution, London, 1988.
- 11 Frijters, J.E.R., “Sensory difference testing and the measurement of sensory discriminability”, in Piggot, J.R. (Ed.), *Sensory Analysis of Foods*, 2nd ed., Elsevier Applied Science, London, 1988, pp. 131-54.
- 12 O’Mahony, M., “Short-cut signal detection measures for sensory analysis”, *Journal of Food Science*, Vol. 44, 1979, pp. 302-03.

## Further reading

British Standards Institution (BSI), *BS 5929: Methods for Sensory Analysis of Foods Parts 3 and 8*, British Standards Institution, London, 1984, 1992.