
Statistics for food science III: sensory evaluation data. Part A – sensory data types and significance testing

John A. Bower

The author

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

Abstract

Introduces statistical methods employed in analysing sensory data. Describes significance testing and simple procedures for determination of population characteristics in sensory data and highlights sources of error and replication in the sensory context. Discusses factors important in selection of an appropriate statistical test for sensory data.

Introduction

Sensory evaluation is a basic tool in the practice of food science and is widely used in research and industry. The data generated come in many forms and, unlike those of chemical analysis measures, discussed in "Statistics for food science (SFS) II", are produced by judgements of the human subjects who form a taste panel. Although it is becoming possible to mimic the sense of smell instrumentally [1], it is unlikely that such devices will completely replace human panelists. Much importance can attach to the conclusions drawn from sensory experiments. For example, in product development the decision to launch a product with a modified formulation may be based on experimental evidence which has determined that the difference will not be noticed by consumers. If this conclusion proves false and the product is rejected, then high financial losses could occur. Thus the reliability of the methods, and the statistical analyses which decide these issues, are important. Some sensory methods may appear simple on the surface, but professional procedures are highly dependent on good experimental design. This latter topic will be dealt with in more detail in a later article. It is not intended here to give a discourse on sensory methodology as such, but rather to give an appreciation of how statistical analysis plays a part in the control and decision making. Current advances in this field reveal a complex picture, as many factors have a bearing on how people perceive quality and decide preference in foods.

To enable an adequate description of the application of statistical methods in sensory evaluation, it becomes necessary to cover additional statistical concepts and procedures. These could apply equally well to any experimental situation in food science, but sensory evaluation provides an ideal setting for these explanations.

Significance testing

A common task of the sensory scientist is to establish whether or not products or samples differ to a sufficiently large extent so as to be deemed significantly different. This can be in terms of preference or of the intensity of some attribute, such as sourness, spiciness, etc. Another possibility is to assess individual panellists for significant performance of a

sensory ability. Significance testing involves a standard procedure by which hypotheses are formulated and tested under specified conditions. The tests commence by assuming a “no difference” condition – the *null hypothesis* – and the proposition of an *alternative hypothesis* which will be accepted if the null hypothesis is rejected. An experiment is carried out to test the evidence, as obtained from a sample, against the null hypothesis, and a statistic is calculated. This may be in the form of a probability value or a test statistic which is compared with a critical value from tables which give the chance probabilities based on a true null hypothesis.

Two outcomes are possible: if the probability of the result is sufficiently low or the test statistic exceeds the critical value, then the null hypothesis is rejected, the alternative hypothesis accepted and a significant result is concluded. Otherwise the null hypothesis is retained: it cannot be proved true as such; rather it is concluded that the data did not provide sufficient evidence to warrant declaring it false. The term “significance level” refers to the probability level at which the test is operated. Similar in concept to the confidence level (SFS II), the conventional, most conservative significance level is 5 per cent. Obtaining such a significance would mean that there is confidence that similar samples drawn from the population will show such differences 95 per cent of the time. Thus there is a possibility, admittedly small, that the experimental results were purely due to chance alone. As can be seen, there are *risks* associated with significance testing – one is the risk of wrongly concluding a significant result which is in fact absent (known as a “type I error”); the other is the risk of not concluding a significant result which is actually present (a “type II error”). Although interdependent, these risks depend also on the magnitude of the difference between the samples – larger differences will be easier to detect and therefore will reduce the risks. Both risks can be reduced by increasing sample size but, as with confidence intervals, achieving zero for both so as to be 100 per cent certain could be impractical, as we are again making population inferences from a sample. The probability levels for these risks are referred to as *alpha* (α) and *beta* (β), respectively. An important associated term is the *power* of the statistical test, a measure of the test’s ability to find a true significance,

where power equals $1 - \beta$. The lower the β risk, the more powerful the test. The choice of significance levels is arbitrary: traditionally 5, 1 and 0.1 percentages are commonly used, but there are circumstances when a less stringent α such as 10 per cent may be justifiable – for example, in experiments with low sample and panellist numbers, used to assess rough trends.

A source of much confusion in significance testing is the use of one- and two-tailed tests. This refers to whether or not the test considers a *directional* difference in the samples examined. The “tails” can be visualized as the areas under the distribution frequency curve (described in SFS II) which lie on either side of the central region, although the region of difference could lie within one tail only and perhaps one- and two-sided is a better description. If the experimenter can predict the direction of difference, assuming the null hypothesis is false, then a one-tailed test would be appropriate. For example, in a triangle test the panel organizer knows that out of three samples – A, B, B – A is the odd one out. If the null hypothesis (which states “There is no detectable difference”) is false, then it can be predicted that the panellists will select A. In a paired preference test where two formulations, X and Y, are compared, it is unlikely that the outcome could be predicted if there is a detectable difference in preference and a two-tailed test is used. It is “easier” to get a significant result with a one-tailed test because of lower chance probabilities, but it must be fully justified if selected. If there is doubt, or if the sensory analyst wishes to be conservative, then a two-tailed test is the option. In fact, statisticians themselves differ on this point, some advocating only two-tailed tests on the basis that the experimenter can never be sure regarding the outcome. This is particularly relevant to sensory evaluation where the decision-making process through which individual panellists go is usually unknown to the panel organizer.

The *degrees of freedom* of a significance test refer to the number of terms which are independent in the calculation of the test statistic. Thus, if there are n terms in the raw data, calculation of a mean uses up one degree of freedom. Subsequent statistics calculated using the mean will have $n-1$ degrees of freedom. In addition to this use in calculation of statistics, the number of degrees of freedom is required for locating the critical values from

tables. Degrees of freedom obviously are directly related to sample size and also have a bearing on the way in which inference from sample to population is made (SFS I). In certain types of experiment, full assessment of statistical measures may not be possible unless there are sufficient degrees of freedom.

Some of the main factors which influence sensory data can now be considered. The food scientist presented with sensory evaluation results or engaged in operation of sensory panels should consider the following points.

What kind of factors can affect the reliability of sensory data?

Sensory evaluation is subject to error sources just as is any measurement system, although the error terms of SFS II are not used to the same degree in the sensory context. Many influences can operate to affect precision and accuracy and hence possibly reduce reliability.

A number of factors, mostly *psychological* in nature, arise because of the use of human subjects as measurement tools. The panel organizer will issue instructions to panellists; yet, even with training, confusion can arise regarding the semantics of these instructions, as well as the use of the chosen sensory scale. Thus, individual panellists may differ in their interpretation of sensory attributes, the intensity level and the general decision-making process involved, etc. In fact, the decision criteria may even change within the test[2], further complicating the issue. Tasting acuity can vary and a “warm-up” effect can occur where initial discrimination is poorer[3]. One or more panel members may lack concentration, be easily distracted, or suffer from fatigue or some health condition which affects tasting ability. Even resting and palate rinsing, etc., intended to introduce control over fatigue, can lead to loss of memory regarding the sensory impressions[4]. Panellists can pre-decide sensory impressions because of expectations induced by verbal and written instructions and by visual and odour cues. Cross-sensory confusion can occur if the panellist must judge one aspect of sensory properties when receiving multiple impressions or taints from unintended sources. Other factors include appetite level, preferential use of some numbers on certain scales, and the fact that perception of intensity can depend on the intensity range within the sample set.

Several *experimental design* considerations can have significant effects on reliability of

data. The presentation of samples is crucial and must be standardized with regard to form, temperature, conditions of lighting, non-associative coding, etc. [5]. Presentation order is well known for its bias-causing effects, and random or balanced order is necessary. Taste sensations from initial samples can linger on and modify subsequent perception. This “carry-over” effect can be reduced at a physical level by palate rinsing; but the mental impression is imprinted by the first or preceding sample, thus enforcing the requirement for balanced presentation order[6]. Judging strength or intensity of a sensory attribute may be haphazard unless calibrated by control or reference samples.

.....
 ‘...Expert panels are used in laboratory situations...’

The importance of the above factors varies according to the type of panel, the level of training and the purpose of the sensory experiment. Expert panels are used in laboratory situations. They are usually required to be able to detect small differences in sensory attributes and be able to gauge intensities with accuracy and consistency – such panels can be viewed as instruments, and any factor which affects these abilities becomes important and may require control to minimize its influence. The operation of panels with less training, consumer or otherwise, is viewed differently. The purpose is to gain information on characteristics which are detected by consumers or to assess attitudes and preferences. If used under laboratory conditions, then influencing factors must be controlled; but the degree of control may be less or unnecessary for the particular method. Balanced against this is the view that such laboratory control leads to an unrealistic representation of consumer perception[7]. Home-use testing would seem to answer this problem, but it is the most variable and uncontrolled method. If control of factors is exercised, then it is required to cover variability across assessors, within sessions and across different sessions.

Although panel training and experimental control can reduce the errors, they have the potential to “fog the issue”, leading to a weakening of power or to the creation of apparent significant effects which are in fact false. Certain statistical techniques can “even-out” some error sources, such as the different use of scale by panellists[8].

How can the error level be measured for sensory data?

The *precision* (variability) within replicates for sensory data can be measured by the same techniques as are used for chemical and instrumental data – mean deviation, standard deviation, etc.; others are possible, such as the scoring ranges used by Seaman *et al.* [9] in which the use of correlation coefficients was also employed to compare large numbers of duplicates. Confidence intervals are less used in sensory work, although the importance of this statistical measure has been emphasized [10, 11]. The magnitude of precision obtainable by sensory panels can be quite acceptable, especially for expert panels, but is unlikely to compare with those of some modern analytical instruments. Panel and panellist precision can be monitored and maintained by training. Depending on the manner in which replicates are presented, coefficients of variation of 5–10 per cent or better should be attainable.

Accuracy features less, in general, but adequate control and calibration are essential for some applications. Accuracy is perhaps more applicable with trained expert panels, where the panel is employed as an instrument which is able to perform analytically. Control samples or reference chemicals are used to provide panellists with base-line measures against which to gauge test samples. These calibration samples can be for odours, tastes, colours and textures. Single or multiple standards can be used to provide reference for one or more points on the intended scale [12]. Some of these standards have international status, as in the case of certain colour standards. In difference tests where there is a correct result, a measure of accuracy is given by the level of correct responses for a panel or for individual panellists. Selection of panel members can be based on their accuracy with difference tests and ranking of graduated concentrations of tastes, etc. Meilgaard *et al.* [13] suggest 40–60 per cent accuracy for a triangle test, depending on the difficulty of the test.

What level of replication is required in sensory evaluation?

Replication in sensory evaluation differs from the circumstances of chemical or instrumental analysis because there is more than one measuring “unit”, in the form of the individual panellists. Thus, even with only one tasting

per judge, a replicate measure is obtained. True replication requires each panellist to perform a judgement two or more times on samples of the same product. The type of sensory method, the panel type and the purpose of the particular experiment will decide whether or not replication is included. With trained panels, many replicates may be used to gauge panellist precision, namely, four to six for descriptive profiling and ten or more for difference testing. Replication is lower in or absent from consumer panels, but panel size is larger. The numbers of product samples examined are limited because of taster fatigue; thus from four to eight tastings are typical, although with visual judgements more can be accommodated. In cases where the number of samples with or without replicates becomes too large for the panel, it is possible to use experimental design-blocking techniques to reduce the work. Recommended panel size depends on the sensory method and its purpose; trained panels having lower numbers. As before (see SFS II), replication, although desirable in principle from the statistical standpoint, is limited by *cost* in terms of time and personnel.

Bearing in mind the error sources above, replication in sensory testing may be “self-destructive” to some degree. In an attempt to gain a more confident result, a panel organizer may decide to include replicates, but because of limitations of panel availability these are included in the same session. This increases the number of judgements to be made, and sensitivity may decrease through fatigue or a panellist may recognize or remember a previous replicate and be biased. Ideally, replicates should not be influenced by previous members of a set – hence the desirability of blocking at separate replicate sessions.

What is the nature of the data generated by different sensory methods?

The sensory methods are well established and have been detailed by Seaman *et al.* [14]. The types of data which they generate are shown in Table I. It is pertinent to bear in mind what these methods are used to assess – that is, acceptance testing or detection of differences, etc. and the type of panel used, whether expert, trained or consumer.

Table I also shows whether or not the population from which the data come is assumed to be normal. These points are

Table I Properties of data generated by sensory methods

Sensory method or scale	Type of data generated	Assumptions regarding data
Difference test	Nominal	Non-normal
Ranking	Ordinal	Non-normal
Scaling, scoring	Ordinal, interval	Possibly normal
Magnitude estimation	Ratio	Normal

important because statistical tests are based on certain assumptions which, if not true, can invalidate the tests. The level of measurement, the nature of the population distribution from which the data come and the equality of variance for treatments (product samples) examined are major points. *Parametric tests* assume that the data (or in some cases statistically-derived values) are normally distributed, at the least interval and continuous in form and have equal variances for treatments. *Non-parametric tests*, which are less powerful, are “distribution-free” in this respect and make less rigorous assumptions regarding population parameters.

.....
 ‘...The issues of most indecision lie with scaling and scoring methods...’

In some cases, there may be doubt concerning these conditions, and the sensory analyst faces the task of deciding the issue. The issues of most indecision lie with scaling and scoring methods. These can be integer point scales (e.g. 9- or 7-point hedonic scales) or graphic line scales with end anchors plus possible intermediate anchors. It has been shown[15] that some panellists are reluctant to use the extremes of such scales, resulting in unequal intervals (“end-effects”). The imposed limits to the scale may cause “squeezing” of the distributions of test samples which have mean scores near the extremes – there is insufficient room for both tails of the distribution to spread out symmetrically. This causes deviation from normality and inequality of variance.

How can the data be assessed for normality and equality of variance?

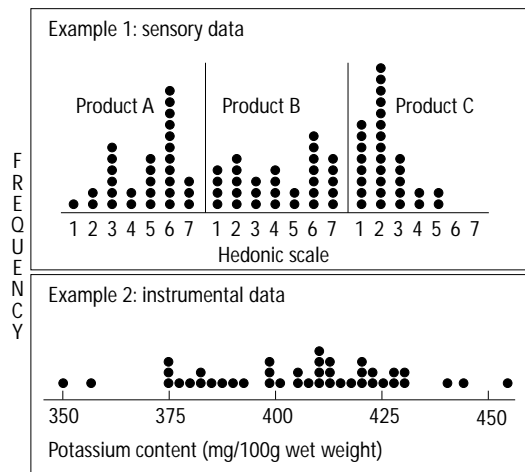
Testing data for normality can be done by simple graphical techniques. Figure 1 shows blob charts or dotplots of results from experiments with relatively small data sets which illustrate some common features. The perfect bell shape or Gaussian curve of the normal

distribution is not apparent in the examples and even with larger samples may still be absent. Look for symmetry on either side of the plot and bunching or indication of a peak in the central region. Products A and C of the sensory example exhibit peaks, but they are not centrally located. Product B is more symmetrical, but there is a suggestion of two peaks. These results were obtained using untrained consumers. The instrumental data (obtained by atomic absorption spectroscopy analysis during a study on potassium levels in potatoes[16]) show a peak more centrally located and there is more overall symmetry.

Relatively simple analysis will yield other measures (Table II). These include *skewness*, a measure of symmetry, and *kurtosis*, a measure of the “peakedness” or “flatness” and also of unimodality or bimodality in a distribution. These measures, which are based on deviations from the mean raised to the powers three and four, respectively, are laborious to calculate manually and are attainable using a software package[17] for the data of Table II. Another measure of skewness, detailed by Rees[18], is included for its simplicity of calculation:

$$\text{Skewness} = 3 \times \frac{(\text{sample mean} - \text{sample median})}{\text{sample standard deviation}}$$

Figure 1 Dotplots of distribution of sensory and instrumental data



Formulae for the more complicated coefficients of skewness and kurtosis are given in [19,20].

Skewness is zero when symmetry is perfect; negative and positive values occur when a tail of the distribution is extended to the left or right respectively. Kurtosis is zero for a normal distribution; negative and positive values indicating bimodality and unimodality, respectively. Additionally, for perfect symmetry, the mean, median and mode must be equal.

Product C (sensory) and the instrumental data show apparent symmetry by these latter characteristics, but this is not supported by the high magnitude of the skew for C. Products C and A show high positive and negative skew, respectively – a case of squeezing of the distribution caused by the nature of the scale and the location on it. The instrumental data also show the lowest kurtosis, but product B exhibits strong evidence of bimodality although it has more symmetry. Bimodality can signify disparity among panellists in choice of categories on the scale, unimodality being more desirable.

Even samples taken from a truly normal population are likely to exhibit non-normal values for the measures above. So how does one decide what degree of deviation from normality is acceptable? A check on skew and kurtosis based on the characteristics of the normal population is described by Smith[19]:

$$\begin{aligned} &\text{Extent of skewness} \\ &= \text{skewness} \times \sqrt{\frac{\text{No. of samples}}{6}} \end{aligned}$$

$$\begin{aligned} &\text{Extent of kurtosis} \\ &= \text{kurtosis} \times \sqrt{\frac{\text{No. of samples}}{24}} \end{aligned}$$

If the absolute magnitude of either of these measures exceeds 1.96, then it can be concluded that the data come from a population which is not normal. Using these criteria, it can be seen (Table II) that only the data of product C fail in this respect, although overall the sensory data exhibit greater deviation from perfect normality. The simpler skew measure [18] must exceed absolute unity in order to be considered marked; in this case none of the experimental data show this level of skew. Small sample sizes (ten or less) would yield less useful information on application of

Table II Distribution measures of sensory and instrumental data

Measure	Data			Instrumental
	A	B	C	
Mean	4.68	4.20	2.23	405.21
Median	5	4	2	409
Mode	6	6	2	409
Standard deviation	1.68	2.14	1.14	23.04
Variance	2.81	4.58	1.29	–
Skewness ^a	-0.60	-0.14	+1.02	-0.25
Skewness ^b	-0.42	+0.28	+0.62	-0.49
Kurtosis ^a	-0.76	-1.47	+0.67	-0.06
Extent of sk.	-1.33	-0.32	+2.29	-0.67
Extent of ku.	-0.85	-1.64	+0.075	+0.08

Notes: ^a See [17] ^b See [18]

these measures. Trained panels are more likely to use a sensory scale with equal intervals, and usually hedonic scales do produce rough normality unless the product is liked well or disliked well resulting in the skewed nature above [20].

There are other tests for normality, such as the *normal probability plot*, which again is more conveniently done using a statistical software package. Here the experimental values (grouped or single) are plotted against the expected values from a normal distribution source. A linear relationship is obtained if the observed data are from a normal population. Normal plots, prepared using software [17], are shown (Figures 2 and 3) for the potassium data and for product C of the sensory data; the more linear nature of the instrumental data is clear.

Another assumption for some statistical tests is that different treatments compared by the method have equal variances. Equality of

Figure 2 Normal plot of instrumental data (potassium content)

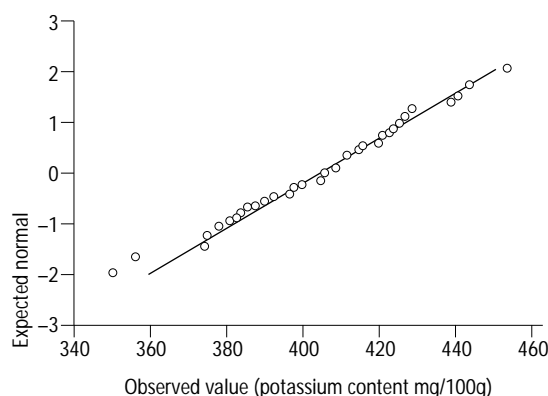
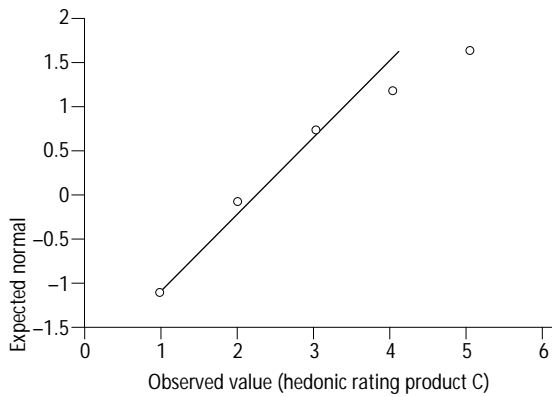


Figure 3 Normal plot for sensory data (product C)



variance can be checked by a simple visual examination of the calculated values (Table II) or by a variance ratio test, where:

$$F \text{ (variance ratio)} = \frac{\text{larger variance}}{\text{smaller variance}}$$

Similar variances will have smaller *F* ratios. This test statistic is compared with tabulated values for the *F* distribution which enables the decision to be based on a significance test. The variance values for the sensory products differ (Table II), but an *F* test shows that only in the case of C compared with the other variances is the difference sufficiently large to be significant. If such inequality was obtained for treatments within one experiment, then one of the assumptions for parametric analysis would be false.

Equality of variance check for sensory data (F test):

- (A cf. B) *F* = 1.63 (not significant)
- (A cf. C) *F* = 2.18 (significant at 5% level)
- (B cf. C) *F* = 3.5 (significant at 1% level)

What can be done if the data are not normally distributed?

It should be borne in mind that the assumptions for normality need only be approximate [18] and that for larger samples such deviations become less important. Additionally some parametric tests are quoted as being robust or resistant to deviations from normality, and can be used anyway [21]. Another possibility is that if non-normality is detected it may be possible to transform the data to produce normality, but this makes the task of interpreting the results more complex [20].

The sensory scientist may have to make the final decision after examining the data as

above. The conservative approach may result in loss of information by not using the more powerful parametric methods, but an incorrect assumption of normality could lead to invalid conclusions. Some practitioners [15,22] recommend doing both types of analysis. If both methods agree, then the conclusion is clear; if not, then a decision guided by experience will have to be made, or more experimentation carried out. Lyon *et al.* [23] recommend use of non-parametric methods for all consumer data and parametric methods for data produced by trained panels with established training records.

(Selection and description of appropriate statistical tests for sensory methods follow in Part B.)

References

- 1 Corcoran, P., "Electronic odour sensing systems", *Electronics & Communication Engineering Journal*, October 1993, pp. 303-8.
- 2 Ennis, D.M., "Relative power of difference testing methods in sensory evaluation", *Food Technology*, Vol. 44 No. 4, 1990, pp. 115-17.
- 3 O'Mahony, M., Thieme, U. and Goldstein, L.R., "The warm-up effect as a means of increasing the discriminability of sensory difference tests", *Journal of Food Science*, Vol. 53 No. 6, 1988, pp. 1848-50.
- 4 Matthews, M.A., Ishii, R., Anderson, M.M. and O'Mahony, M., "Dependence of wine sensory attributes on vine water status", *Journal of Science and Food Agriculture*, Vol. 51, 1990, pp. 321-35.
- 5 Williams, A.A. and Arnold, G.M., "The influence of presentation factors on the sensory assessment of beverages", *Food Quality and Preference*, Vol. 3, 1991/2, pp. 101-7.
- 6 Macfie, H.J. and Bratchell, N., "Designs to balance the effect of order of presentation and first order carry-over effects in hall tests", *Journal of Sensory Studies*, Vol. 4, 1989, pp. 129-48.
- 7 Lawless, H.T. and Claassen, M.R., "Application of the central dogma in sensory evaluation", *Food Technology*, Vol. 47 No. 6, 1993, pp. 139-46.
- 8 Naes, T., "Handling individual differences between assessors in sensory profiling", *Food Quality and Preference*, Vol. 2, 1990, pp. 187-99.
- 9 Seaman, C.E.A., Hughes, A.H., Hinks, C.E. and Parry, D.A., "Consumers as sensory panellists", *British Food Journal*, Vol. 95 No. 8, 1993, pp. 7-8.
- 10 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.
- 11 Smith, G.L., "Statistical properties of simple sensory difference tests: confidence limits and significance tests", *Journal of Science and Food Agriculture*, Vol. 32, 1981, pp. 513-20.

- 12 Ishii, R. and O'Mahony, M., "Use of multiple standards to define sensory characteristics for descriptive analysis: aspects of concept formation", *Journal of Food Science*, Vol. 56 No. 3, 1991, pp. 838-42.
- 13 Meilgaard, M., Civille, G.V. and Carr, B.T., *Sensory Evaluation Techniques*, Vols I and II, CRC Press, Boca Raton, FL, 1987.
- 14 Seaman, C.E.A., Hughes, A.H., Hinks, C.E. and Parry, D. A., "How is sensory quality measured?", *Nutrition & Food Science*, No. 5, September-October 1993, pp. 15-19.
- 15 O'Mahony, M., "Some assumptions and difficulties with common statistics for sensory analysis", *Food Technology*, Vol. 36 No. 11, 1982, pp. 75-82.
- 16 Bower, J.A., "Cooking for restricted potassium diets in dietary treatment of renal patients", *Journal of Human Nutrition and Dietetics*, Vol. 2, 1989, pp. 31-8.
- 17 SPSS Inc., *SPSS for Windows, Release 6.0*, SPSS Inc. Chicago, IL, 1993.
- 18 Rees, D.G., *Essential Statistics*, 3rd ed., Chapman & Hall, London, 1995.
- 19 Smith, G.L., "Statistical analysis of sensory data", in Piggot, J.R. (Ed.), *Sensory Analysis of Foods*, 2nd ed., Elsevier Applied Science, London, 1988, pp. 335-79.
- 20 Gacula, M.C. and Singh, J., *Statistical Methods in Food and Consumer Research*, Academic Press, Orlando, FL, 1984, pp. 23-60.
- 21 Land, D.G. and Shepherd, R., "Scaling and ranking methods", in Piggot, J.R. (Ed.), *Sensory Analysis of Foods*, 2nd ed., Elsevier Applied Science, London, 1988, pp. 155-86.
- 22 Vie, A., Gulli, D. and O'Mahony, M., "Alternative hedonic measures", *Journal of Food Science*, Vol. 56 No. 1, 1991, pp. 1-5, 46.
- 23 Lyon, D.H., Francombe, M.A., Hasdell, T.A. and Lawson, K., *Guidelines for Sensory Analysis in Food Product Development and Quality Control*, Chapman & Hall, London, 1992.

Further reading

- Amerine, M.A., Pangborn, R.M. and Roessler, E.B., *Principles of Sensory Evaluation of Food*, Academic Press, New York, NY, 1965.
- O'Mahony, M., *Sensory Evaluation of Food: Statistical Methods and Procedures*, Marcel Dekker, New York, NY, 1986.
- Stone, H. and Sidel, J.L., *Sensory Evaluation Practices*, Academic Press, London, 1985.