
Statistics for food science IV: two-sample tests

John A. Bower

The author

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

Abstract

Describes statistical methods applied to two-sample significance tests. Discusses independent, related and paired samples, Wilcoxon, sign, Mann-Whitney U and *t*-tests and illustrates their application.

In "Statistics for food science (SFS) III" *Nutrition & Food Science*, Vol. 96 No. 2, the emphasis was on sensory data. The series could continue in this vein, but rather than giving the impression that the statistical methods or tests can apply only to one form of data it is appropriate to include all data forms likely to be generated by food scientists. A common form of rapid experimentation can be performed by comparing two treatments or a control with one test treatment. Data from simple experiments of this nature can be analysed by a number of different techniques. Some of the important features of these tests are covered below but the reader should consult statistical texts for fuller details of underlying assumptions and specific limits to sample numbers, etc.

What are "two sample" tests?

The term "two samples" refers to two groups or sets of data which are compared in respect of some measured feature, e.g. a chemical constituent or a physical parameter, or which are rated or scored by sensory means for some quality aspect. The question is often one of deciding whether the two groups differ significantly or are differences simply due to sampling variation? The analysis of such data would depend on the points regarding scale type, population distribution form, etc. and whether or not the samples are related. There are many statistical tests for comparing such data and a number of the more common ones are detailed in Table I. The binomial and chi-square tests (SFS IIIb) can also be applied to this situation to establish if the proportional balance between two groups (nominal data) is significantly above chance probability in nature. Certain of these tests can also be used as one-sample tests, as when comparing a single sample result with a population parameter.

Other assumptions for these tests include: a continuous nature to the variable or the variable differences; equality of variance (independent *t*-test); and a random nature and independence of sampling, etc. If a related samples test is used then the data must be matched as pairs. The continuous nature of the variable may not be apparent on the scale used, but if the scale could be structured (theoretically if not practically), so that a measurement could take any fractional value, then this is adequate. The tests are applicable with small sample numbers (2-30

Table I Statistical tests for comparing two groups of data

Test	Sample nature	Measurement scale	Minimal assumptions	
			Population distribution(s)	Population parameters tested/compared
Sign test	Related	Ordinal	None	Median of differences
Wilcoxon signed rank	Related	Ordinal	Symmetrical	Median of differences
Mann-Whitney U test	Independent	Ordinal	Shape/symmetry	Median of each group
Two-sample <i>t</i> -test	Independent	Interval	Normal	Mean of each group
Paired <i>t</i> -test	Related	Interval	Normal	Mean of differences

approximately depending on the test); for larger sample sizes, other tests, usually based on approximations to a normal distribution, are recommended. Some tests are parametric in nature, others non-parametric and there are examples for related and independent samples. This latter distinction can cause difficulties and is worthy of further explanation.

Are the two sample groups related?

The correct decision regarding this point is easy to misjudge and can be a common cause of applying the wrong statistical test. One key issue lies in the source of the data and the relative degree of independence existing between and within the sample groups. This consideration is important for any experimental design, whether two, or three or more sample groups are involved.

Independent sample groups

In cases where there is no cross relationship between individual values from one group with those of the other, then the samples are said to be independent. A common feature of such data sets is that a series of true replicates is generated within each group. For example, analysing two cultivars of fruit for ascorbic acid content, using one method of analysis on five replicate samples from a single fruit of each. Here, the samples in each group are obviously originating from a separate independent source. A less obvious example would be found when applying two analytical methods repeatedly to one material to determine whether there was any bias between the methods.

In the case of sensory evaluation an independent test is usually applied when comparing data from one panel with that of a second panel, possibly on samples of the same product. Scores for each sample (between sample) are coming from different groups of people. A

rarer sensory example of an independent design type would be given when a single assessor scores replicate samples of two products. Here the within-sample score independence is assumed to be sufficiently low to justify classifying the design as independent.

The statistical methods applied to such groups are often referred to as two-sample tests, where the intention is often to test hypotheses regarding the population means or medians of the two groups. The two groups do not need to be the same size in terms of sample numbers for these tests.

Related or paired sample groups

If individual samples in one group are linked or related in some way to the corresponding samples in the second group, then the samples are classed as related or paired. Certain extraneous variation will be eliminated and a different form of analysis is applicable. Any differences arising in the pairs on application of treatments should arise solely due to the treatment and not because of original differences being present. Statistical analysis is done using paired tests and the intention is to examine the difference between the pairs of values, and to test hypotheses regarding the mean or median of the population of differences. Differences between means or medians usually require to be less for similar significance when compared with those required by an independent test.

In many cases of paired measurements there may not be replication within the groups – e.g. comparing two methods of analysis on a series of different food materials where only one determination per method is performed on each material (assuming any errors are independent of analyte concentration). Each pair of values is unique to the source material. The within sample independence is now greater.

A single sensory panel evaluating two treatments is usually a related design. Each assessor

presents a pair of scores; any independence between samples is masked by the greater degree of independence within the groups owing to use of different assessors. Any differences detected would be due to the different sample treatments and not because different assessors were judging within each group. “Before and after” experiments on the same material also result in a paired design (“self-pairing”), as in comparing a single panel’s ability before and after training.

Pairing and matching

With foods it is often possible to divide one item into two (or more) parts to achieve pairing. The division should be such that the resulting halves are similar in chemical composition and structure. Similar matching can be done with food lots in respect of size, storage age or any other characteristic which is known to be similar. It should not be assumed that the use of such samples automatically ensures a related design – if the degree of within sample independence is sufficiently low an independent design is more appropriate. The experimenter must ensure that a true paired relationship exists across the sample groups.

As stated above, comparing two different groups of people is usually considered to be an independent design, but matching can be done with human subjects also. The ultimate match could possibly be obtained by the improbable example of pairs of cloned individuals with the same psychological profile. Twins, followed by pairs of siblings are lesser, more likely possibilities. Such pairs are not numerous and it is more practical to attempt to achieve pairing by matching individuals on features such as age, sex, occupation, etc. Characteristics related to the response to be measured should be included, e.g. liking or disliking a particular food or being a “regular purchaser of” a food product. In some disciplines applying these criteria could provide adequate assurance of pairing, but they are unlikely to be regarded as sufficient in sensory evaluation studies. There is an increasing risk of individual attitudes affecting the results, thus weakening the case for a related design. If the matching is poor or based on a characteristic which is not truly related to the response variable then true differences will be more difficult to detect. The term “matching” can also be applied in a looser sense to the standardization of experimental conditions apply-

ing to sample groups. This ensures that any possible further sources of variation are minimized or eliminated, but does not necessarily imply the paired relationship.

Which type of design is better – two-sample or paired?

Selection of design type before experimentation is usually more important than attempting to select the correct analysis method after the fact, so the question often arises as to which type of design is “better”? Usually a related design is preferable because of the fact that it should allow a more focused examination of the treatment effect and possibly detect smaller differences. Also it may reveal a wider amount of information as in the case of applying two analytical methods to a range of different foods rather than replicates of one. Some problems may not be answerable by use of a paired design as in the case where it is not possible to present all test samples to a single sensory panel at the one time and two different panels are employed.

Which type of two-sample or paired test is appropriate?

Parametric tests

When normality assumptions hold, then t -tests are appropriate. t -tests are available for related and independent samples, and even for cases where the assumption of equality of variance in the parent populations is false. Such tests would usually be applicable to chemical and instrumental data and also to sensory data where the assumptions hold. The test can be used for any number of samples over two within each group, although greater than five is desirable; with over 30 samples the t -distribution and the normal distribution become so similar that another test (the z -test) based on the latter distribution can be used instead. The test compares the difference of the sample means (independent) or the mean of the differences (related) as modified by a measure of how the data vary. The variation measure is given by an estimate of the sampling distribution population variance known as the standard error of the mean. The estimate is calculated using a single or pooled sample or difference standard deviation and the square root of the sample number. A similar estimate was used in SFS II (*Nutrition & Food Science*, Vol. 95 No. 5) to determine a

confidence interval. The calculated statistic is compared with tabular values which give the maximum *t*-value, which could occur by chance for a true null hypothesis. For significance at the stated level, the calculated statistic must lie outwith the “acceptance region” limits which are defined by the tabular values. Usually this means that calculated *t* must equal or exceed the tabular *t* in absolute magnitude, although equality may depend on the degree of accuracy (number of decimal places) with which tabular values are given. Statistical software packages often give both highly accurate tabular values and also the probability of getting a value as extreme, or more extreme than the observed *t*: the “*p*-value”. If the *p*-value is less than or equal to α then the null hypothesis is rejected.

Non-parametric tests

Non-parametric tests are used if the above assumptions are deemed invalid or in cases with small sample numbers and possible outliers in the data. If the design is paired in nature then tests such as the Wilcoxon signed rank test (matched pairs) and the sign test can be applied. The latter test reduces the data to a simple difference in sign between pairs and thus the magnitude of difference no longer applies and statistical power is lost. This can be acceptable, especially where the experimenter has doubts concerning the measurement system and the scale used, and there is a possibility that the data are basically ranked in pairs. For each pair of data, one value is subtracted from the other, but only the sign of the result is recorded, ties being ignored. The significance of the incidence of positive and negative signs is assessed by binomial probability, i.e. if the null hypothesis is true then positive and negative signs would appear with equal incidence. A large incidence of ties (>25 per cent of the data values) would lessen confidence in the sign test result.

The Wilcoxon signed rank test assumes symmetry in the population of differences. The statistic is also calculated by ranking, but in this case as applied to the signed numerical differences between pairs. The smallest difference is given rank one, tied ranks are each given a mean rank and differences of zero are discarded. Thus the magnitude of the differences exert an effect in this test. The signed ranks which occur least are summed and compared with the critical value for the circumstances of the test. Occurrence of 25 per cent or more of ties

lessens the accuracy of the critical values of this test and in such cases an approximation to normality test is recommended[1]. The Mann-Whitney U test assumes similarity of distribution shape for both groups and can be used as an alternative to the independent samples *t*-test. The analysis is done on a ranked set of all values from both data groups.

The incidence of ties can affect the above non-parametric tests and this can be a problem in the sensory context. The use of a limited category or discrete scale offers more chance for scores or measures to be the same. A more continuous scale such as a graphic line analogue type would give a higher probability of the occurrence of differences between pairs, especially if numerical anchors are omitted or positioned at the extremes.

Illustration of the tests

The computation and formulae for the above tests are given in most statistical texts, but software packages offer a more rapid comparison of several examples. Consider the paired data sets (Table II) and their analysis by several of the tests (Table III). The data could represent instrumental (experiment 1) and sensory (experiment 2) measures. Although the figures differ between the experiments, they produce the same results on application of statistical tests. The mean and median of the two groups are different but variances are sufficiently similar (an *F*-test is non-significant), although this is not a requirement for a paired test.

It can be seen that the probability level of significance obtained increases with the power of the tests. The sign test is unable to

Table II Example data sets for paired samples

	Experiment 1		Experiment 2		Difference (A-B)
	Group A	Group B	Group A	Group B	
	59.0	56.0	8	5	+3.0
	60.0	58.0	9	7	+2.0
	55.0	53.0	4	2	+2.0
	57.0	55.0	6	4	+2.0
	59.0	56.0	8	5	+3.0
	53.0	55.0	2	4	-2.0
	58.0	55.0	7	4	+3.0
Mean	57.286	55.429	6.286	4.429	1.857
Median	58.0	55.0	7.0	4.0	2.0
SD	2.498	1.512	2.498	1.512	1.773
Variance	6.238	2.286	6.238	2.286	3.143

Note: Some statistical values are rounded off for simplification

Table III Analysis of paired data (Table II) by different statistical tests

Test ($\alpha = 0.05$)	<i>p</i> -value (two-tailed) ^a
<i>Correct application</i>	
Sign test	0.125
Wilcoxon matched-pairs signed-ranks test	0.052
Paired <i>t</i> -test	0.032
<i>Incorrect application</i>	
Mann-Whitney U test	0.128
<i>t</i> -test for independent samples	0.118

Note:
^a Values rounded off for simplification

achieve a rejection of the null hypothesis, even though there is a preponderance of positive signs. This illustrates a feature of the binomial test with small sample numbers – for a two-tailed test with $p < 0.05$ all seven signs must concur. The Wilcoxon test achieves $p = 0.052$, and only when the more powerful *t*-test is applied does $p < 0.05$ appear. Assuming that the paired designs are valid, then incorrect application of independent tests results in a different overall conclusion. Non-significant *p* values are obtained – for the small differences exhibited by these data the independent *t*-test has been unable to establish significance.

It should be noted that the above result pattern is not necessarily typical for data in general. If the differences between group

means and medians are large then independent tests may achieve higher levels of significance than paired tests, again assuming that one form of the tests is incorrectly applied. The potential of the non-parametric tests should not be discounted bearing in mind the points raised in SFS IIIa. The sign test, although the weakest test, can compare favourably with more powerful tests, depending on sample size and the nature of the population distribution[2]. If departures from normality are extreme then a sign test could provide a more valid analysis than a *t*-test.

References

- 1 Mahony, M., *Sensory Evaluation of Food: Statistical Methods and Procedures*, Marcel Dekker Inc., New York, NY, 1986, pp. 304-9.
- 2 Gacula, M.C. and Singh, J., *Statistical Methods in Food and Consumer Research*, Academic Press, Orlando, FL, 1984, pp. 323-7.

Further reading

- Bender, F.E., Douglass, L.W. and Kramer, A. (Eds), *Statistics for Food and Agriculture*, Food Products Press, New York, NY, 1988.
- Miller, J.C. and Miller, J.N., *Statistics for Analytical Chemistry*, 3rd ed., Ellis-Horwood, Chichester, 1993.
- Rees, D.G., *Essential Statistics*, 3rd ed., Chapman & Hall, London, 1995.