

---

# Statistics for food science V: comparison of many groups (part A)

---

*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 three or more sample groups. Discusses analysis of variance in parametric forms and the requirement for experimental design control before its application.

Experiments which compare three or more treatment groups or conditions cannot usually be analysed using the methods described in "Statistics for food science (SFS) IV" where two group comparisons were detailed. Such experiments demand a wider, more versatile technique which forms the analysis basis of many simple to complex experimental designs. The first question concerns possible incorrect application of two-sample and paired tests to the many sample group situation.

## Can two-sample or paired tests be applied to many sample comparisons?

In practice it is possible to perform this procedure – the data set from the many-sample experiment is split into pairs, each group paired with each other one. A series of the appropriate independent or related sample tests can then be used, but from a statistical standpoint this approach is not recommended, and is usually condemned. There is an increased risk of committing a Type I error (SFS IIIa) and rejecting a null hypothesis which is in fact true, i.e. accepting a result as significant when it occurred by chance. This arises because of the increased number of comparisons being performed and applies to any "multiple testing" circumstance. A statistical test done with a single pair, at the 5 per cent level involves a 5 per cent risk of committing a Type I error. This condition applies to randomly selected samples and subsequent randomly selected pairs which are independent of any other pair. For experiments which have three or more sample groups, pairing would result in many more comparisons, e.g. five treatment groups demand ten inter-comparisons, not all independent of one another, and these would be at an increased risk level; for seven groups 21 tests would be required and one of these would commit the error. The procedure may be justified if the experiment was organized in a paired or two-sample manner, but this constitutes a different type of design which would usually require a different form of analysis. The separate treatment "global" approach has organizational advantages. It is simpler and easier to present a series of single treatments for analysis or assessment than to prepare many sets of pairs, especially for sensory evaluation experimentation in food science.

### Which analysis method is suitable for comparing more than two sample groups?

For such experiments the recommended statistical design and analysis approach is by analysis of variance (ANOVA). Irrespective of treatment numbers the risk level is maintained at the specified significance level. Some of the simpler designs where ANOVA is applied are given in Table I. Terminology of these designs and tests can vary slightly, depending on individual texts. Other important tests include tests based on chi-square which can be applied to analysis of categorical data from three or more groups.

Although experiments using these techniques can appear to be straightforward in operation, the design features become increasingly important and form the foundation of many comparative experiments.

### What does experimental design entail?

Experimental design should operate in all tests, trials and experiments within any discipline where control is required. In its simplest form, it is the adoption of an organized manner for planning, execution and analysis of the work. Initial design steps would be to establish clear objectives for the experiment and to postulate hypotheses to be tested. In the context of ANOVA, design would include decisions on which factors are being examined, what levels they are to be set at and the methodology to be employed for measurement of the response. A typical experimental procedure might begin with division of the food material(s), which is to be experimented on, into individual lots or batches, the size of

which depends on the scale of the experiment, ranging from laboratory scale to full production scale. These experimental units are randomly allocated to the treatments which are the specific settings or combinations of factor level(s). Any source of variation can be examined as an experimental factor. A process condition such as temperature, or an ingredient quantity in a food formulation would be examples of quantitative factors where the variable is continuous. The levels of these factors would be the individual temperatures or the ingredient weights respectively, i.e. different magnitudes of the factor. Factors can also be qualitative, and set at two or more levels where magnitude is not considered, e.g. the type or identity of an ingredient such as the type of salt in a formulation. Here the levels could be sodium chloride and potassium chloride. Other possible sources of variation could include sensory panellists or panels, individual analysts, instruments, process machines, laboratories, etc. In certain experiments involving human subjects as samples, the treatment factors would be in the form of the different conditions under which the groups of subjects were examined. All settings for these factors are under control and are the independent variables. The response is the dependent variable which is measured at the end of the experiment on each treatment. One or more responses can be measured, e.g. the concentration of a chemical constituent, or the magnitude of a physical, instrumental or sensory parameter.

The simplest experiment would involve a minimum of one factor at two levels but beyond this there is theoretically no limit. Obviously the more factors and levels which are included, the larger the experiment

Table I Some experimental designs analysed by ANOVA (comparison of three or more groups of data)

Design/test	Sample nature	Some assumptions			Parameter tested/ compared
		Measurement scale	Population distribution(s)		
<b>Parametric ANOVA:</b>					
Completely randomized design (CRD) 1-way	Independent	Interval	Normal		Means of groups
Randomized block design (RBD) 2-way	Related	Interval	Normal		Treatment effect
<b>Non-parametric:</b>					
Kruskal-Wallis 1-way ANOVA (CRD)	Independent	Ordinal	Identical except for location		Distribution; median
Friedman's 2-way ANOVA (RBD)	Related	Ordinal	None		Distribution; central tendency (median)

becomes. Consideration must be given to all other sources of variation and uncontrolled factors which might affect validity. If possible, these must be controlled or limited using design techniques such as randomization, blocking and replication. Alternatively, there are statistical procedures which can circumvent these problems or allow some quantification of them so that they can be taken into account in eventual analysis. The importance of replication has been discussed several times in this series and blocking is dealt with below where design types are examined. Randomization has been proven to reduce error and bias caused by taking a systematic, non-random approach to organization of experimental stages. The technique involves randomization of the order in which stages and individual treatments are performed or allocated, e. g. in sensory experiments designs to counteract order and other bias effects such as carry-over should be employed, etc.

Adequate controls or base line measures should be included so that the test results can be gauged relative to these. A control treatment is often represented by an experimental unit which is actually "untreated". In an experiment studying the effect of processing conditions the control could be a unit consisting of raw unprocessed food. Space precludes giving a fuller account of experimental design at present. Some features of an experiment in food science research[1] which employed a number of these design features is illustrated in Table II.

The control for this experiment was a formulation which excluded the dairy ingredients (the main factor of interest). A large number of experimental runs were required to accommodate the many factors and levels causing the researchers to chose partial replication, albeit of randomly allocated full treatments.

As with other statistical tests there are both parametric and non-parametric forms of ANOVA (Table I). Much of the explanation below applies to parametric analysis of variance. Parametric ANOVA can also be applied to two data set comparisons in which case it is equivalent to a *t*-test, but the commonest application is for three or more groups.

### What does ANOVA do? – the ANOVA method

The design details above indicate that sources of variation are assigned under various factors and levels. The technique of ANOVA partitions this source of total variation into its component parts, enabling an assessment of the magnitude of effect of each source on the response(s). Essentially ANOVA compares variability between treatments with that within treatments. The between treatment variation is caused by the nature of the treatments, e.g. the difference in an analytical, instrumental or sensory measure caused by different food processes or formulations. The within treatment variation is due to the variation in the values for the replicates within each treatment group, caused by several error effects. This error effect is also referred to as the residual variation, i.e. the "left over" variation which cannot be accounted for by treatment or factor effects. The variance (SFS I) from these two sources is used to calculate two different estimates of population variance. These measures, referred to in ANOVA terminology as mean squares (MS), are calculated for these data groups and their relationship is expressed as a variance ratio (*F*) :

$$F = \frac{\text{MS of treatments}}{\text{MS of error}}$$

Table II Example of design features of an experiment analysed by ANOVA

Factor	Type	No. of levels	Level identity
Dairy ingredient identity	Qualitative	5	Na caseinate, skim-milk, Na caseinate (high viscosity), Whey protein, Demineralized whey
Starch concentration	Quantitative	2	2%, 4%
Process temperature	Quantitative	2	76°C, 82°C
Experimental units:	5kg batches of smoked meat sausage		
Response measures:	Sensory properties (15 attributes), instrumental texture (Instron testing instrument), instrumental colour (Minolta chromometer)		

Source:[1]

The  $F$  ratio for the experiment is compared with tabular  $F$  values, calculated for different levels of significance over a range of experimental conditions, i.e. the numbers of treatments and replicates, expressed as the degrees of freedom. If the experimental  $F$  ratio exceeds the critical  $F$  value (usually at the 5 per cent level or less) then a statistically significant result is obtained for the data as a whole. If the null hypothesis is true, i.e. there is no evidence to suggest that the population means are not equal, then within the context of the experiment, the treatments are not demonstrating a measurable effect. In this case the two estimates of population variance (via the treatment measures and the background error) will be similar and the  $F$  ratio will be low – theoretically equal to unity.

If the null hypothesis is false and there is a real and detectable effect due to the treatments, then depending on the design and precision of the experiment, the treatment variance will be greater in magnitude than the error variance. Thus, it can be seen that a maximum for  $F$  is obtained when the variability between treatments is high combined with low variability within treatments (i.e. good agreement for replicates). If there is high variability within replicates then  $F$  falls; similarly  $F$  falls if there is low variability between treatments, i.e. the treatments may not be demonstrating sufficient difference for significance or detection. An insignificant  $F$  ratio does not prove that the treatments are having no effect – it may be that the experiment was too small to detect the differences. A larger experiment would possibly achieve this, but the practical significance of such a smaller difference would have to be scrutinized – the scientific or commercial importance must be considered.

### What is the difference between within and between treatment variance?

It is important to clarify the exact nature of replication which gives rise to the within treatment variation. It is possible with some forms of design to apply each treatment once and to perform a single response measurement on each. This procedure would be limited in the extent of analysis due to insufficient degrees of freedom. Additionally, within treatment variation cannot be determined without replication. Another approach is to apply each treatment once and draw replicate samples for end analysis. For such a sample

set there is likely to be more than one source of variation: one is due to random error at the measurement stage – end measurement variation or analytical error; the other is caused by sampling variation due to inherent variation within the food material. The magnitude of this latter variation will depend on the food, e.g. a series of experimental units drawn from freshly mixed whole milk will be more homogeneous in chemical composition than those drawn from a day's harvest of root vegetables. Whether or not both these sources of variation are measured and hence make an impact on the statistical analysis, depends on procedure. If a treatment is applied once and several samples are taken for response measurement, then the variability of this set of results usually reflects the analytical variance. Any inherent variation effect will depend on the form of the experimental unit at the point of end assessment. If the unit is made more homogeneous, by the treatment itself or by the end analysis method, any inherent variability is likely to be lessened. Inherent variation and "unit to unit" variation can be more appropriately estimated by comparison with a repeat treatment on a similar experimental unit. The data (Table III) for an analytical experiment [2] illustrate this. Replicate end analyses were performed on duplicate samples drawn from macerated single experimental units of raw tissue from different potato tubers. The variability, expressed as the percentage coefficient of variation (%CV; SFS I) is compared with that between several experimental units from the same source.

The treatment per cent CV is much greater due to the relatively large differences in potassium content between different lots of raw material. Thus, with raw material of this nature, even before application of treatments the experimental units may differ considerably and these differences have the potential to swamp any treatment effect. In cases of a single application of treatments, such circumstances may give a false impression of the experiment's precision. Apparent significant effects between treatments could be due to high inherent variability. Additionally, there is a danger in cases where only one full treatment is performed: it must be truly representative, i.e. if it were done again would a similar result be obtained? If a gross error occurs it may not be detected but a duplicate treatment could show up the error. To overcome such

problems the experimenter must ensure adequate sample size and some degree of full treatment repetition within the experiment.

In sensory experiments this procedure is often not followed and it is common practice to draw multiple servings from single product batches for panellist assessment. Although more work and planning would be involved, it is recommended[3] that full replication of each production run is included in the design.

**What assumptions does ANOVA make?**

The assumptions for this analysis method are similar to those of the *t*-test (SFS IV) in terms of normality of distribution and equality of variance between groups, etc. If the same measurement system is used on each group (same method, same analyst or same sensory panel, etc.) then the homogeneity of variance is a reasonable assumption. Normality of distribution is assumed mainly in the error term, with a random, independent or uncorrelated nature to the error. The treatment effect is assumed to be independent of level – this can be infringed in analytical experiments where the variance may depend on the concentration of the analyte.

A more complex assumption lies in that of additivity of factor effects. At a fundamental level the ANOVA method commences by specification of a model which describes, in mathematical terms, the effect that treatments have on the response measure in the experimental units – the ANOVA model. Essentially, ANOVA assumes that as treatments are applied to the experimental units the eventual result on the response will be modification of the population mean by a linear addition of these effects plus error. The specific magnitudes of the effects (or coefficients) of the model can be derived from the experimental data by the process of ANOVA.

**What if the ANOVA assumptions are broken and how can the data be assessed?**

As stated previously (SFS IIIa) such assumptions need only be approximate, and procedures such as transformation can be applied to allow data to “qualify” in this respect[4] . The experimenter is recommended to establish some awareness of the status of the data, especially as they may reveal additional information – spurious data, data entry errors, hidden trends, etc. Visual examination can often show trends if the data are of small size, e.g. inconsistency between replicates, obvious differences in treatment means and even suspicion of non-additivity. Suggested checks would include inspection by graphical means, to allow an impression of distribution shape and characteristics (if sample size permits), as well as indication of any outliers. Performance of some analysis on features such as homogeneity of variance and normality of distribution is useful. A measure of agreement within replicates via standard deviation or coefficient of variation can quickly ascertain whether this variability is within expected ranges. Measures which reveal the uncertainty of the sample estimate, such as confidence intervals and standard errors, are also recommended. After ANOVA, diagnostic checking by examination of residuals in terms of distribution and sequence can reveal additional information on adherence to assumptions. Significance tests are available for several of these procedures to aid in decision making.

In practice, research publications in food science rarely include all of the above checks in the published material, but such procedures may improve publication potential[5]. Modern software packages allow rapid performance of many such techniques.

**Which form of ANOVA and experimental design is applicable?**

If doubt exists regarding the above assumptions then the results obtained by parametric

Table III Potassium content (mg/100g wet basis) of raw potato for replicate end analyses (atomic absorption spectroscopy) and replicate treatments (raw controls)

Experimental unit (500g)	Replicate end analysis (5g)	Mean	% CV	
Replicate treatment 1	426,422	424	0.67	Average %CV for end analysis = 0.93
2	389,383	386	1.10	
3	409,415	412	1.03	

Notes: Overall mean = 407

Average % CV for repeated full treatments = 4.77

ANOVA can only be viewed as indicative of differences, and non-parametric ANOVA should be used if an appropriate design is available, otherwise the parametric form (which is more powerful) is applicable. Other important points include the general design features of the experiment in respect of whether the samples are related or independent (SFS IV). This depends on the experimental objectives and any limits which are imposed and needs to be decided at the planning stage.

If independent sample groups are warranted then a completely randomized design (CRD) is applicable with each experimental unit being randomly allocated to a group. In its simplest form this design deals with variation of one factor and the analysis method is referred to as “1-way ANOVA”. It is commonly applied to simple experiments in food science, involving chemical, physical or instrumental measure changes caused by factors such as processing, etc. For effective use the experimental units must be homogeneous or “equally variable” within limits of inherent variation. Food experiments where the response measurement is by objective or instrumental techniques can often meet this requirement. In the case of sensory experiments, the CRD would require separate groups of panellists for each treatment and they would be assumed to be uniform in their ratings[6]. This is less likely and limits the use of the CRD for much sensory work. Such a design could apply in a consumer study where each consumer was only able to be tested once, or possibly in storage studies where the original panellists were not available for each assessment stage during storage. The problem is avoided by use of a related group design where the panellist effect is viewed as a second factor. If a CRD is used in sensory studies it is recommended[6] that in order to counteract panellist variation a large number (> 100) of consumers is recruited. The CRD does not require equal numbers within each treatment group which could be a useful feature in the above sensory studies.

The related groups circumstance of a single sensory panel assessing several treatments requires a block design such as the randomized complete block (RCB), with data items in each treatment group having corresponding members in each other group. The variation in the blocks could be considered as another factor but is usually removed so as to

give more focused attention on the main treatment factor. Thus, on analysis the procedure for “2-way ANOVA” is applicable. The block effect can be any factor which is likely to have a potentially uncontrolled effect on the response – e.g. if the experiments require to be done over several days, then each day could constitute a block. The blocks could also be a series of different process machines, different laboratories, analysts or indeed, as above, panellists. In all these cases the variation due to different days, machines, or panellists, etc. would be calculated and removed from the error term. An improved estimate of treatment effects is obtained and the RCB design is one of the commonest design types, possibly because of its relatively simple manner of increasing experimental precision.

Beyond these simple designs the number of factors which can be examined is theoretically unlimited but practical considerations limit most experiments to a selected few. Thus there are 3-way, 4-way designs, etc., the complexity of calculation increasing with the number of factors. There are distinct advantages in examining the effect of several factors together rather than one at a time. This approach is ultimately more economical of experimental time and resources. In addition to an assessment of main effects, such designs which include replication allow detection of interaction between factors. The presence of significant interaction indicates that factors are not operating independently but depend on the levels of other factors. More complex designs are available to cover a wide variety of experimental demands[7,8]. For non-parametric analysis the options are more limited, and for the more complex designs there may be no readily available non-parametric equivalent.

The analysis method and the information obtained in all these designs also depends on the exact nature of the treatments. If these are under direct control and are selected by the experimenter they are described as fixed effects and subsequent inferences only apply to those factors and levels. A random-effects model, with random selection of factors and levels, allows inference to a wider population of possible factors and levels. The ANOVA calculation differs for these two cases and looks at means and variances respectively. A mixed-effects model is also possible with both fixed and random factors. The fixed-effect model is the more common. Random effects

would apply to a sensory study where a randomly selected panel of consumers was enlisted to measure product preference. The random-effects model would allow inferences to be made about the consumer population's preference attitudes.

Irrespective of the number of factors, replicates, etc. one or more  $F$  ratio values are obtained, each stemming from a source of variation (factor effects). If the  $F$  ratio is sufficiently large and its  $P$  value is equal to or less than the stated  $\alpha$ , then a significant effect has been detected for the experiment as a whole.

### Does ANOVA identify individual treatment differences?

A significant variance ratio indicates that at least two of the treatments differ in respect of the hypothesis being tested. It does not indicate where the significant difference(s) lie and a pair-wise comparison must be performed. This will be detailed in part B along with illustrations of parametric ANOVA and data evaluation procedures.

### References

- 1 Baardseth, P., Naes, T., Mielnik, J., Skrede, G., Holland, S. and Eide, O., "Dairy ingredient effects on sausage sensory properties studied by principal component analysis", *Journal of Food Science*, Vol. 57 No. 4, 1992, pp. 822-8.
- 2 Bower, J. A., "Cooking for restricted potassium diets in dietary treatment of renal patients", *Journal of Human Nutrition and Diet.*, Vol. 2, 1989, pp. 31-8.
- 3 Meilgaard, M., Civille, G. V. and Carr, B.T., *Sensory Evaluation Techniques*, 2nd ed., CRC Press, Boca Raton, FL, 1991, pp. 257-62.
- 4 Steel, R. G. D. and Torrie, J. H., *Principles and Procedures of Statistics – a Biometrical Approach*, 2nd ed. McGraw-Hill International, Singapore, 1981, pp.167-71.
- 5 Huck, S. W. and Cormier, W. H., *Reading Statistics and Research*, 2nd ed., HarperCollins College Publishers, New York, NY, 1996.
- 6 Gacula, M. C., *Design and Analysis of Sensory Optimization*, Food & Nutrition Press, Trumbull, CT, 1993, pp. 29-34.
- 7 Box, G. E. P., Hunter, W. G. and Hunter, J.S., *Statistics for Experimenters*, John Wiley & Sons, New York, NY, 1978.
- 8 Cochran, W. G. and Cox, G. M., *Experimental Designs*, 2nd ed., John Wiley & Sons, New York, NY, 1957.

### Further reading

- Bender, F.E., Douglass, L.W. and Kramer, A. (Eds), *Statistics for Food and Agriculture*, Food Products Press, New York, NY, 1988.
- Chatfield, C., *Statistics for Technology*, 3rd ed., Chapman & Hall, London, 1992.
- Cohen, S.S., *Practical Statistics*, Edward Arnold, Sevenoaks, 1988.
- Gacula, M.C. and Singh, J., *Statistical Methods in Food and Consumer Research*, Academic Press, Orlando, FL, 1984.
- Miller, J.C. and Miller, J.N., *Statistics for Analytical Chemistry*, 3rd ed., Ellis Horwood, Chichester, 1993.
- O'Mahony, M., *Sensory Evaluation of Food – Statistical Methods and Procedures*, Marcel Dekker Inc., New York, NY, 1986.
- 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.