British Journal of Nutrition (2001), **86**, 141–144
DOI: 10.1079/BJN2001409

Review article

Statistical analysis of nutritional studies

Graham W. Horgan

Biomathematics and Statistics Scotland, Rowett Research Institute, Aberdeen AB21 9SB, UK

(Received 20 February 2001 – Revised 24 April 2001 – Accepted 30 April 2001)

Statistical analysis of data collected in experimental or observational studies is an important part of nutritional research. If it is not done appropriately, there is a risk that information in the data is lost, or that conclusions are misleading. The present article does not attempt a full review of the subject of statistics, but attempts to list aspects where care is needed.

Statistics: Data analysis: Randomisation: Experimental power

Nutritional studies draw on many scientific disciplines, and one that must be prominent in any list is statistics, the science of variability. The consideration of variation is something which is inescapable in any nutritional research undertaking, whether it be experimental or observational, on human subjects or animals, or based on laboratory cultures. There is always a need to describe variation, and to draw conclusions in the presence of variation.

Many scientists find an understanding of statistics doesn't come easily, however great the need. Some of this lack of understanding is due to statistics having too often been presented as a branch of mathematics, which it is not (mathematics is one of its tools). Some of it is also due to the abstraction of the subject's central ideas, such as the need to consider not only the variation in experimental material, but also the uncertainty in sample summaries and in conclusions drawn.

I do not intend to give a review of, or introduction to, the subject as a whole. There are many textbooks suitable for this purpose (and, regrettably, many which are not!). Instead, I will draw attention to ten issues in the statistical treatment of nutritional studies which are often misunderstood or confused. I expect that most statisticians working in nutrition would devise a fairly similar list. A considerable influence on the selection presented is the reviewing of many papers submitted for publication in the *British Journal of Nutrition*.

The order of the list is not necessarily intended to reflect importance. The first six points are general, the next three relate specifically to experimental studies, and the tenth relates to observational or survey studies, where much less control, or none, of the effects and mechanisms being investigated is possible.

It may not help to categorise

It seems to be a human instinct to lump information into groups. This solution is not always helpful when done with measurements which show continuous variability. It is common, for example, with different levels of intake of nutrients or exposure to hazards to group individuals into (typically) between two and five ranges of intake or exposure, and then to base statistical analyses on differences between groups. This approach appears to be adopted mainly for ease of presentation. It means that one can quote differences between groups, which are easy to understand. We can, for example, talk about individuals being lean or obese as if these were distinct groups with different characteristics, even though the truth is that very many individuals are close to the cut-off BMI of 25 kg/m^2. The drawback of categorising is that we throw away information. How much will depend on the distribution, but will typically be about 36 % for two groups, 21 % for three groups, 14 % for four groups and 11 % for five groups (based on a normal distribution and linearity of effects).

A lesson may be that if we must group, then we should have five or more groups. One use of such grouping is to detect non-linearity of response to the grouping variable.

Standard deviations and standard errors are not interchangeable

The standard deviation, standard error and standard error of difference are all widely used indicators of variability, although sometimes rather indiscriminately. They are all useful, but they provide information about different aspects of variability.

Corresponding author: Dr Graham Horgan, fax +44 1224 716608, email g.horgan@bioss.ac.uk

The standard deviation indicates how variable measurements are. It should be quoted if this aspect of variability is of interest; for example, if we wish to compare variability in two populations.

The standard error indicates how precise an estimate is. It can be obtained for any quantity estimated from a sample, such as a mean, a proportion or a regression slope. It should usually accompany any estimated quantity quoted in a paper, where these estimates are of interest in themselves, rather than compared with some other estimate. An example might be the proportion in a population suffering some medical condition, or a regression coefficient relating a response and explanatory variable.

The standard error of difference is the standard error of the difference between two estimated quantities, usually two means. It is quoted when interest is in such a difference; for example, as an estimate of a treatment effect. The standard error of difference is more often required than the standard error, reflecting the greater usefulness and interpretability of comparisons than simple estimates, in nutritional studies.

A 95 % CI is a range of values consistent with the quantity being estimated. If calculated correctly, and based on valid assumptions, 95 % of such intervals will contain the true but unknown value of the quantity being estimated.

Any of these indicators could be used to draw error bars on a chart, but it should be stated explicitly which is being used. The current convention seems to be to use the standard error. If comparisons are the main interest, then the standard error of difference or CI are more appropriate.

Don't revere the 5 % significance level

A *P* value is a measure of evidence. The smaller the value, the greater the evidence against a simpler model (e.g. no effect, no association) than the one of potential interest (an effect or an association). The use of $P=0{\cdot}05$ as a cut-off is a convention which has an historic basis rather than a scientific, mathematical or philosophical basis. Papers have been submitted to the British Journal of Nutrition with *P* values such as 0·045 and 0·055 used to draw opposite conclusions for the comparisons to which they referred. This is nonsense; the *P* values are almost the same. While few individuals would quarrel that *P* values $>0{\cdot}1$ indicate negligible evidence, and those $<0{\cdot}01$ indicate considerable evidence, anything between these values should be discussed with caution rather than confidence.

It may be confusing that, being measures of evidence rather than providers of infallible conclusions, results based on *P* values do not necessarily follow the rules of logic. We may have three treatments (A, B and C) and conclude that A and B do not differ significantly, nor do B and C, but that A and C do. This conclusion is logically impossible, if interpreted rigidly. A more appropriate conclusion is that A and C appear to differ, but that the data do not allow us reliably to discriminate among the possibilities that B is similar to A, that it is similar to C, or that it is somewhere in between.

Does the power of the experiment justify the claims?

It is often quoted that 'absence of evidence is not the same as evidence of absence', but it seems frequently to be forgotten in the writing up of experimental work. A *P* value $>0{\cdot}05$ is often used to justify a conclusion that no effect or association exists. The most formally correct way to address this issue is in terms of power calculations; the probability that the experiment would detect as significant an effect of a specified size, if it existed. Put simply, how likely are we to detect an effect if it exists? This likelihood might be rather lower than the experimenter imagines.

Another way to consider the extent of evidence against the existence of an effect is to examine a CI (95 % or 99 %) for the size of an effect. If this value is, for example, 'mean 10 (CI −2, +22)' (in whatever units are appropriate), then the data are quite consistent with the effect being zero. However, they are equally consistent with it being 20. If one chooses to dwell on the former in discussing the result, then this preference is based on knowledge or experience rather than the data. There may be some merit in replacing a bland 'no significant effect' conclusion with something like 'it is unlikely that the effect exceeds 22'.

Non-parametric tests lack power

Many researchers appear not to realise that non-parametric tests (such as the Mann–Whitney, Kruskal–Wallis, Wilcoxon etc.) are less likely to detect effects and associations than more standard distribution-based approaches (such as *t* tests, ANOVA etc.). This is largely because the tests discard some of the information in the data. Most tests are based on rank orderings only, rather than the absolute values of the measurements. They have the disadvantages that they are unrelated to the data summaries (means and standard deviations) which will also necessarily be reported, and versions for even moderately complex experimental structures are hard to find.

Authors often opt for non-parametric tests because of concerns about normality, or because their use is thought to show greater caution. Normality concerns are often unjustified. Some experimenters have unrealistically high expectations of how close the histogram of a small data set should resemble a perfect Gaussian curve. Where distributions are clearly skewed, then a transform, such as the logarithm, should first be considered. Non-parametric tests should be used only as a last resort, usually with data which have many zeros (e.g. numbers of cigarettes smoked per d), or only take a small number of distinct values (such as number of lambs born to a ewe, or observations based on a five-point subjective scoring system).

Multiple comparison tests are usually best avoided

The idea behind multiple comparison tests is appealing. If we are testing several factors at once, such as comparing two or more treatment regimens with a control, or comparing each of the treatments with each other, then the overall risk of a statistical test giving a false positive is greater than the *P* value at which each test is done. Multiple comparison tests are a way of adjusting for this factor, and getting bigger *P* values as a result. They have the one advantage of dampening excitement about just-significant *P* values. However, on the whole they more often hinder

rather than help interpretation. There are many different tests (Tukey, Duncan's, Student–Newman–Keuls, Bonferroni, Dunnett's, Hsu's and many more), all taking different views of the comparisons to be made. Few scientists or statisticians will be able to recall what each test is aiming to do. All tests have somewhere the idea of a set of comparisons of equal interest, which is rarely true in practice. It is usually best just to present simple *P* values and remind the reader that type I error rates (false positives) apply to each test, not to them all taken together. This point is more important when the comparisons being tested were suggested by examination of the data rather than having been specified in advance as part of the experimental protocol. If one is determined to use a multiple comparison test, then a brief explanation of what it does, and a reference to the literature, should be provided.

Blocking is not just for horticulturalists

It is perhaps unfortunate that one of the most powerful devices in experimental design, blocking, is a term which derives from its historical development in crop research. It refers to the arrangement of experimental material (animals, human subjects, or whatever units experimental measurements are taken on) into groups such that variation within groups is less than for the material as a whole. If treatments are applied within groups, then the random variation against which treatment effects are assessed will be less than if the grouping had not been done. Managing the experiment in groups can simplify planning, and ensure that any extraneous sources of variation (differences over time, between individual experimenters etc.) end up as differences between groups, where they will not affect the precision of assessments of treatment effects. Often there is very little extra cost and effort in such grouping, and much to be gained. Any measurement which can be taken before the experiment starts, and which is likely to be correlated with the experimental measurements of interest, can be a base for grouping (blocking). It can also be based on the time sequence in which laboratory measurements are made (i.e. all one block first, then all the next block etc.), or the physical layout of media which have to be left to develop for a time in a controlled environment (such as an incubator). The only situation where grouping may be not a good idea is where group differences and sample sizes are small.

Not all comparisons are made on the same level

A common situation in experiments is illustrated by the following example: animals are assigned to different treatment groups in which they remain for the duration of the experiment. Let us use breed as an example (there being no doubt that an animal must remain of the same breed!). In addition, some other factor varies on different occasions for each animal. This factor may be simply the passage of time itself, or it may be something which is done to the animal, such as before and after some treatment, or an alternation of a treatment factor. Let us use diet as an example. The experimenter will be interested in effects of diet, of breed, and whether there is an interaction between them. It may seem natural to examine a standard two-way ANOVA. For the example given earlier, with (for example) two breeds, three diets and six animals of each breed, the ANOVA of the thirty-six observations, three each on twelve animals, would look like:

Source	df
Breed	1
Diet	2
Breed × diet	2
Residual	30
Total	35

where only the columns for the source of variation and df are shown. This ANOVA is wrong. It is mixing a between-animal comparison (breed) and within-animal comparisons (diet and breed × diet). There may be quite different amounts of random variation at these two levels. A correct ANOVA is hierarchical, often termed split-plot (another hangover from the early horticultural days of statistics). For the example given earlier it will look like:

Source	df
Between animal	
Breed	1
Residual	10
Within animal	
Diet	2
Breed × diet	2
Residual	20
Total	35

The between-animal part of the ANOVA shows the sources of variation between animals, including the residual random variation, and the within-animal part shows the sources of variation in measurements at different times on the same animal, again including a residual random part, which will usually be quite different from the between-animal random variation.

Many elementary statistical software programs will not perform ANOVA such as the example given earlier, but for experiments with variation at more than one level it is essential, and programs such as Statistical Analysis System (SAS Institute Inc., Cary, NC, USA), Genstat (NAG Ltd, Oxford, UK), BMDP (Statistical Solutions, Saugus, MA, USA), SPSS (SPSS Inc., Chicago, IL, USA) or any other which will do a correct analysis must be used.

A common source of within- and between- subject or animal observations is repeated measures studies, where the development of a measurement over time is of interest. This approach adds the further complexity of likely non-independence of observations at nearby time points. There are many sophisticated ways of dealing with this factor (for more detail, Diggle *et al.* 1994). A very simple and elegant approach, which avoids most difficulties, is to calculate summaries for each subject (means, contrasts, rates of change) of the sequence of observations and compare these summaries between subjects in different groups. Matthews *et al.* (1990) provides a clear explanation.

Compare whatever received the treatments

Randomisation is essential for ANOVA and *t* tests, which base their calculations on random distributions. Where this

factor is sometimes overlooked is in experiments where measurements are made on different units from those which were randomly allocated to treatments. If an experiment allocates treatments to, for example, sows or female rats, and then takes measurements on the piglets or rat pups, then the analysis should be done at the level of the sow or rat mother, since she received the treatment. The piglets or pups within a litter do not provide independent observations of the treatment effect. They share some genetic relatedness and a common uterine environment. The correct approach is to take averages (or possibly some other summary) over litters, and compare the litter averages.

Nutritional surveys aren't easy

Here we consider a few points relevant specifically to nutritional surveys. The methods of analysis of data collected in such studies tend to be straightforward, i.e. summaries, correlations and tests of differences between subgroups. If there is a problem, it is usually that authors show insufficient awareness of the inherent difficulties in nutritional surveys, which need to be acknowledged when attempting to draw conclusions from results. Any of the standard methods of assessing dietary intake (food diaries, recalls, food-frequency questionnaires) are of doubtful precision or accuracy. Individuals are forgetful and untruthful about what they eat. Studies which have compared different intake measurement methods, sometimes also recording blood chemistry markers, have shown alarmingly low correlations, an indication of substantial measurement error. These problems cannot be avoided, but their effects must be discussed.

Another issue which needs comment in any survey report is how representative those surveyed are of any population. Nutritional differences are known to exist between groupings of individuals based on age, gender, education, socio-economic status, geographical region and many other factors. Those individuals who cooperate with or volunteer for a study may differ from those who don't. Authors must at least speculate on how these aspects might bias their results.

And finally...

Do what you say and say what you do

This eleventh point (breaking my promise to stop at ten!) is to say that it is not enough that proper statistics be done, but also that they must be seen to be done. The statistical methods section of many papers is often very inadequate, making it impossible for the reader to decide whether the authors have treated their data fairly or not. For example, it is not enough to say something like 'data were analysed by ANOVA'. The authors may think it obvious what the details of the analysis must have been, but this may not be so clear to the reader. Full details should be supplied. In the case of ANOVA the outcome measure, the structure used and the factors included should be listed.

Conclusions

The eleven issues discussed are just a series of points which may be useful for anyone planning, analysing or reporting on research in nutritional science. They do not in any way constitute a guide to statistical methods; that would require a textbook, perhaps several. Many textbooks are available, and a bibliography is not attempted here. I am not aware of any textbooks which cover specifically the statistical aspects of nutritional experimentation. Epidemiology is a little better catered for, with books by Frank (1996) and Margetts & Nelson (1997). For experimental studies, many of the standard textbooks, such as Mead (1990) and Cox (1992), while not specifically dealing with nutritional research, are a valuable source of ideas.

I will draw attention to one useful source of statistical understanding, which is the articles published frequently in the *British Medical Journal* by JM Bland and DG Altman. These articles address one topic per article, and explain the ideas with clarity, using examples relevant in medical research. Articles appearing since 1994 are freely accessible at http://www.bmj.org. Their textbooks (Bland, 2000; Bland & Peacock, 2000; Altman, 2001) are also valuable.

Acknowledgements

This work was supported by the Scottish Executive Rural Affairs Department.

References

Altman DG (2001) *Practical Statistics for Medical Research*, 2nd ed. London: Chapman & Hall.

Bland M (2000) *An Introduction to Medical Statistics*, 3rd ed. Oxford: Oxford University Press.

Bland JM & Peacock J (2000) *Statistical Questions in Evidence-based Medicine*, 3rd ed. Oxford: Oxford University Press.

Cox DR (1992) *Planning of Experiments*, London: John Wiley.

Diggle PJ, Liang K-Y & Zeger SL (1994) *Analysis of Longitudinal Data*, Oxford: Clarendon Press.

Frank (1996) *Applied Nutrition Epidemiology*. Gaithersburg, MD, USA: Aspen Publishers.

Margetts BM & Nelson M (editors) (1997) *Design Concepts in Nutritional Epidemiology*, 2nd ed. Oxford: Oxford University Press.

Matthews JNS, Altman DG, Campbell MJ & Royston P (1990) Analysis of serial measurements in medical research. *British Medical Journal* **300**, 230–235.

Mead R (1990) *The Design of Experiments*, Cambridge: Cambridge University Press.