

## EDITORIAL

# Presenting data: can you follow a recipe?

Gordon B Drummond<sup>1</sup> and Brian DM Tom<sup>2</sup>

<sup>1</sup>Department of Anaesthesia and Pain Medicine, University of Edinburgh, Royal Infirmary, Edinburgh, UK, and <sup>2</sup>MRC Biostatistics Unit, Institute of Public Health, University Forvie Site, Cambridge, UK

### Correspondence

Dr Gordon B Drummond,  
Department of Anaesthesia and Pain  
Medicine, University of Edinburgh,  
Royal Infirmary, Edinburgh, 51 Little  
France Crescent, Edinburgh  
EH16 4HA, UK. E-mail:  
g.b.drummond@ed.ac.uk

This article is being published in *The Journal of Physiology*, *Experimental Physiology*, the *British Journal of Pharmacology*, *Advances in Physiology Education*, *Microcirculation*, and *Clinical and Experimental Pharmacology and Physiology*. Gordon Drummond is Senior Statistics Editor for *The Journal of Physiology*.

Brian Tom is in the MRC Biostatistics Unit of the Institute of Public Health in Cambridge, and an affiliated lecturer in the University of Cambridge Statistical Laboratory.

This article is the fifth in a series of articles on best practice in statistical reporting.

'It is exceedingly difficult to explain many statistical concepts in terms that are both technically accurate and easily understood by those with only a cursory knowledge of the topic.'

This wise note appears at the opening of a valuable book on reporting statistics (Lang and Secic, 1997). In this series so far, we have tackled this difficult task, and accommodated the need to 'avoid the fine points and distinctions that would detract from an explanation otherwise adequate for most readers'. In other words, we are writing for a readership of science authors and not for professional statisticians. Even statisticians differ between one another in their preferences and procedures, and for consistency we shall continue to use the book cited above (apart from small deviations) as the basis for a uniform set of suggestions. Ultimately, this will become a substantial list, but we will cross reference this list to the concepts and principles we address in further articles. In this long list of suggestions, we shall give reasons for making these suggestions. A good analogy is a cookery recipe. It helps if you are told why things are done in the way suggested, and the principles (as long as they are sound!) are explained clearly. We can extend this analogy: food writers themselves have differences; the best writers recognize that ingredients differ, and even the infrequent cook knows that precise weights and measures need not guarantee a successful dish.

## Key points

- Consult a statistician about sample size and methods of analysis.
- Describe the statistics exactly in the methods section:
  - exact hypothesis
  - how specifically is this to be tested
  - what assumptions have been made about the data
  - what was studied and how many
- Describe *P* values adequately, beware of 'no difference'.
- Restrain the urge to do multiple tests, and correct for them if you cannot.
- Describe symmetrical samples with a mean, SD and *n* ( $\pm$  is redundant)
- For non-symmetrical samples, either use median and quartiles, or consider transforming the data.
- To indicate precision of mean values, use 95% confidence intervals.
- To express differences between samples, calculate the mean difference and the 95% confidence interval for this difference.
- When comparing groups, provide an estimate of effect size.

We now begin to address the practicalities of how data should be presented, summarized and interpreted. There are no exact rules; indeed there are valid concerns that exact

rules may be inappropriate and too prescriptive. New procedures evolve, and new methods may be needed to deal with new types of data, just as we know that new ingredients may require modified cooking methods. However, most writers should follow reasonable principles based on current practice, although some flexibility is required, as we shall show in this perspective.

Before presenting data, an author should be careful to cover, in the methods section, the actual methods used to carry out the analysis, with the same care and rigour as the other methods used in the research. Just as a knowledgeable scientist should be able to replicate the experiment with the aid of the methods section, a suitably qualified reader should be able to verify the results if given access to the data. In some cases, and probably more often than is the case at present, data analysis may require help from a suitably qualified statistician. Now that requirements for small samples are paramount, statistical expertise is more and more necessary. A single catch-all phrase of a handful of tests, placed at the end of the methods section, is as unhelpful to the reader as it might be to read a short command at the end of a recipe to 'chop, slice, boil, sauté, or bake as necessary'.

Thus, in the methods section, give relevant details of the statistical methods:

- 1 Describe how the results were quantified and the data were analysed. This is not trivial. In many biological experiments, the reports are ambiguous. One is left questioning if the units studied and analysed have been assays, cells, action potentials, offspring or litters. The method used for analysis may need to be justified, particularly if it is unusual.
- 2 State the hypothesis to be tested, and the relevant test variables that will be used to test this hypothesis. This is important, because some tests, such as ANOVA, can yield several terms, and only some of these may be addressing the stated hypothesis.
- 3 Consider the assumptions that have been made to allow data analysis, for example, if the data are symmetrically distributed: plotting the values can be helpful in some cases.
- 4 As we have explained, a *P* value can be more or less important, depending on the context: state what value (i.e. significance level) has been chosen to indicate statistical significance. Although it is rarely done, justification of the value chosen suggests that the author has considered the relevance of this threshold value. State if the test is one- or two-sided. Because the value chosen does not represent a 'magical' threshold, presenting the actual *P* value makes much more sense, particularly when the values are close to the chosen significance level. However, an excess of significant figures for smaller *P* values is unhelpful: the logic of a test is not affected if *P* is 0.0023 or 0.0026. (An exception is genome-wide associations, where *P* values can be very small.) To aid judgement of the importance as well as the strength of any significant value, the effect sizes with confidence intervals are particularly valuable.
- 5 In a previous perspective, we showed how a comparison in a small study may yield a 'non-significant' result. Do not claim, merely because the null hypothesis is not rejected, that the groups compared are equivalent.

Describe how the sample size was determined in order to lessen the impact of any 'Type II' error. To do this requires stating what difference would be considered minimally important. Do not just write 'the groups were not different ( $P > 0.05$ )'. (We discuss the value of a confidence interval below.)

- 6 Repeated comparisons increase the risk of a positive result occurring just by chance, even if there is no effect. Too many comparisons suggest a data-dredging exercise and positive conclusions should be taken with a 'pinch of salt' until further examined in other independent studies. The procedures used to correct for multiple tests vary, and advice may be necessary. A single global statistical test is generally preferable to repeated pairwise testing. Correction for multiple testing is recommended when many *a posteriori* tests are performed, and the method used should be described. These methods vary (Ludbrook, 1998; Curran-Everett, 2000). We will revisit this topic, which is controversial, in a later article. A study design and a test strategy that avoid the need for multiple tests are preferable.
- 7 Choose methods that take account of the correlation between repeated (over time) or related/clustered observations. For example, the method to use could be repeated measures ANOVA.
- 8 Give the name, version and source of the software used to process the data. Some less-specialized software is unsatisfactory.

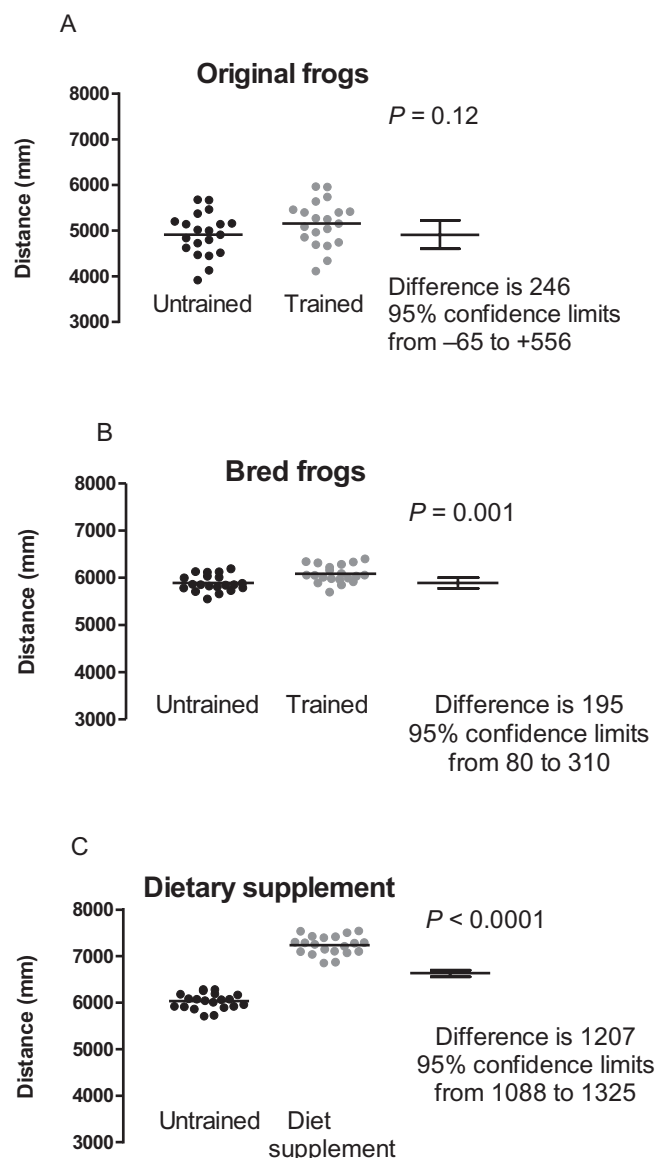
Now let us consider presenting the results. The first article in this series stated 'Show the data, don't conceal them' and this suggestion is frequently ignored. The guidelines for *The Journal of Physiology* currently suggest:

'Data are often better presented graphically than in tables. Graphs that show individual values are better than solid bars indicating a mean value, unless the number of observations is large, in which case a box and whisker plot can be used.' We emphasized two important advantages: the emphasis on central tendency is reduced, and the distribution is explicit. Many modern statistical packages can generate 'dot plots' and all can generate 'box and whisker' plots. It should be possible to understand the figure and caption without recourse to the text.

However, we may need to present some data as numbers. These should be given with an appropriate precision, which is often no more than two significant digits. If data are presented as percentages, then the actual values used for the percentage calculation should be given as well. This is helpful because the denominator may be ambiguous, particularly when a series of values is presented. In the case of percentage change, the change should be calculated as  $100 \times [(final - original)/(original)]$  values.

Most of the time, we wish to summarize a set of measurements and also report a comparison. Let us look at our jumping frogs (Drummond and Tom, 2011). We reported the mean distances that these groups jumped to the nearest millimetre. Perhaps that was a little optimistic, as at the competition the jump length was measured to the nearest  $\frac{1}{4}$  inch!

How shall we summarize the results of that first test on trained and untrained frogs? To describe the first sample we studied (Figure 1) we need to express two different concepts



**Figure 1**

Three comparisons between a control and a treatment group. In panel A, there is no significant difference. In B, the variations of the samples are less, and a smaller difference between the groups is now statistically significant but the difference is small and probably not biologically important. In C, the greater confidence limits are of similar width to those of B but the difference is much greater, and probably important.

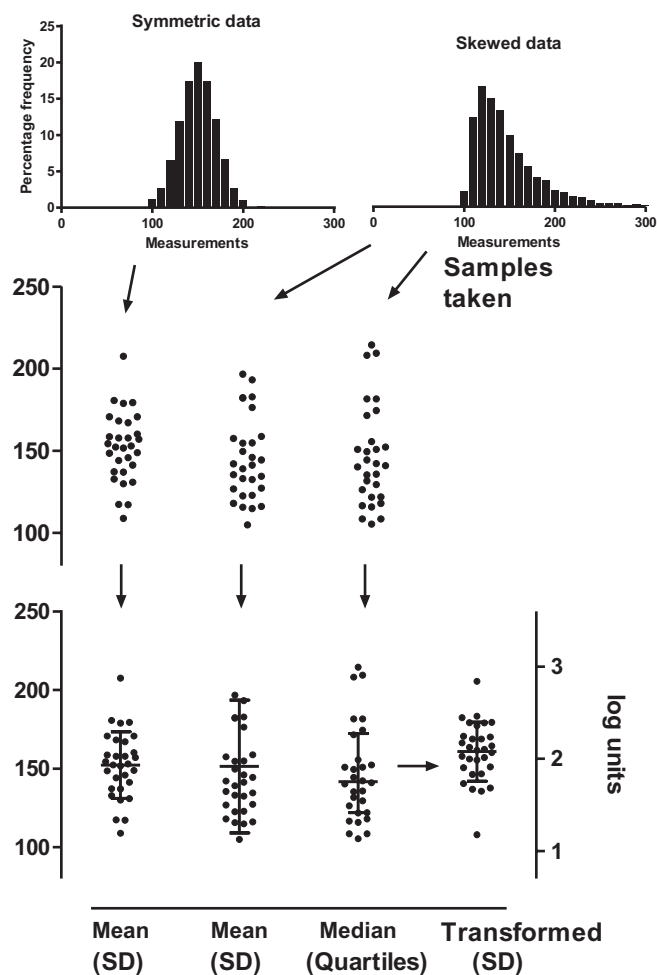
to characterize the samples. We give a measure of *central tendency* to summarize the magnitude of the variable (examples would be the mean or median values) and a measure of *dispersion or variability* (such as a SD or quartiles). In the case of our frogs, the jump lengths and variation of the jump lengths are the important features of our sample. These distances were normally distributed (not a common feature in most biological experiments), so to describe the overall performance of the untrained frogs we report

the sample mean distance jumped. To describe the variability or spread, we report the sample SD. The final value needed to characterize the sample is the number of measurements.

So, in the case of the untrained frogs, we report that the distance jumped was 4912 (473) mm, where these values are mean (SD), and we could add ( $n = 20$ ) if we have not already stated the size of the sample used. These values summarize our estimate of the population characteristics. There is no added benefit in using the symbol  $\pm$  and reporting  $4912 \pm 473$  mm. In fact, the use of this convention can be confusing: is it the mean  $\pm$  SD, the mean  $\pm$  SEM (defined shortly) or a confidence interval?

Many authors choose to use the SEM as a measure of the variability when describing samples. This is incorrect: this value should only be used to indicate the precision with which the mean value has been estimated. As we saw in our frog studies, this value depends very much on the sample size. We told our readers how many frogs we sampled, which is how we achieve the precision. Note that care should be taken when interpreting the SEM as it stands. Here, it is the SD of the sampling distribution of the mean. It tells us how the sample mean varies if repeated samples of the same type (with same sample size) were collected and the mean calculated. In fact, 68% of these estimated means would be expected to lie in a range between one SEM less and one SEM greater than the actual population mean. Thus, there remain a substantial proportion of sample means that do not fall within this range. To assess how precisely a sample mean has been characterized, the preferred measure of precision of a mean estimate is the 95% confidence interval. This allows a reader to more readily contrast mean values. However, to compare formally two mean values, a confidence interval for the *difference* between the means would usually be constructed, as discussed below. Although the relationship between the SEM and SD is straightforwardly related to the number in the sample, it is more considerate of the author to make these calculations and present the reader with a simpler task of comparison.

Most experiments seek to demonstrate an effect, often expressed as a difference between a control group and a group that has been treated. A good way to report such effects is to state not only the mean values for the groups, but also the estimated difference between the measurements, and the confidence limits associated with the difference. Because a common significance level for  $P$  is taken to be 0.05, the common confidence limits used are the 95% intervals. If the study were repeated many times with different samples from the same populations of treated and control frogs, 95% of these range estimates would contain the actual difference between the population means. This confidence interval shows the interplay of two factors, the precision of the measurement and also the variability of the populations, and is an excellent summary of how much trust we can have in the result. The reader can then judge the practical importance of any difference that has been calculated. In Figure 1, which shows our previous frog studies, we can judge the relative importance of training and diet. In panel B, training a less variable population does have a statistically significant result but the effect is small. The impact of diet is also significant, and can also be seen to be much more important. The



**Figure 2**

In the upper panel, we show data samples taken from populations with different distributions. One sample is from a population with symmetrically distributed values; the other two samples are from a population with a skewed distribution. In the lower panel, we can see that the mean values for these groups are very similar, but the SD calculated for the skewed sample is much greater. Using median and quartile values for a sample from a skewed population provides a better description of the data. When the skewed data are transformed into logarithms, the distribution becomes more symmetrical (right-hand axis).

concept of 'effect size' is relevant here and can be expressed in several ways (Nakagawa and Cuthill, 2007). Simply stated in this context, it can be expressed, for example, as the difference between the mean values, in relation to the SD of the groups. However, note that when expressed as a ratio in this way, this measure gives no direct measure of the practical importance of any difference.

Mean and SD are best used to describe data that are approximately symmetrically distributed (often taken to mean normally distributed). Many biological data are not! The shape of the distribution of the data can become evident if they are plotted as individual values as suggested (Figure 2). Another indication of lack of symmetry or a skew

in the distribution (often interpreted as 'non-normality' of the distribution) can be inferred when the SD has been calculated, and this value is found to be large in comparison to the mean.

With a normal distribution, about 95% of the values will lie within 2SD of the mean of the population. For example, we might study a particular type of frog. We find that in a sample the mean distance jumped was 90 cm and the SD of the jump lengths was calculated to be 65 cm. If the distribution was normal, this would imply that approximately 95% of the jump lengths would be within 2SD values greater than and less than the population mean. The lower limit appears in this case to be -40 cm and unless we allow backward jumping, that is not very likely! Although the SD remains a valid estimate of variation (Altman and Bland, 2005), it is less helpful for distributions that are not symmetric, and there are alternative methods for analysis that are perhaps more appropriate.

Non-symmetric distributions can be presented using median and the quartile values. For example, in Figure 2, the 'skewed' sample can be described as having an estimated median of 141 with an interquartile range of (122, 142), where 122 and 142 are the first and third quartile values (the 25 and 75 percentiles). Alternatively, we could transform the data into a form that makes it more symmetric. Values that have been calculated as a ratio, for example as '% control', can be highly skewed. This is a common method of presenting data in many experiments. In such cases, the range of possible results may be limited in the lower values (it may be impossible to obtain values that are less than 0%), but not for the larger values (easy to obtain 150%, or 300%). In such cases, the logarithm of the values may be more convenient for analysis. Rank order tests such as the Wilcoxon do not specifically test for equality of median values, so transforming the data to a more symmetrical distribution may have an advantage. However, when presenting data in a figure, it can be helpful to present in the original scale, as a logarithmic scale is less easy to appreciate (as can be seen in Figure 2).

Although such suggestions have not received universal acceptance, and valid differences of opinion have been voiced, most guidelines advocate these procedures. An easily applied checklist for authors and editors will help their incorporation into practice.

## Conflicts of interest

None declared.

## References

- Altman DG, Bland JM (2005). Standard deviations and standard errors. *Br Med J* 331: 903.
- Curran-Everett D (2000). Multiple comparisons: philosophies and illustrations. *Am J Physiol Regul Integr Comp Physiol* 279: R1-R8.

Drummond GB, Tom BDM (2011). How can we tell if frogs jump further? *Br J Pharmacol* 164: 209–212.

Lang TA, Secic M (1997). A note to the reader. How to Report Statistics in Medicine. Annotated Guidelines for Authors, Editors, and Reviewers. American College of Physicians: Philadelphia, p. xi.

Ludbrook J (1998). Multiple comparison procedures updated. *Clin Exp Pharmacol Physiol* 25: 1032–1037.

Nakagawa S, Cuthill IC (2007). Effect size, confidence interval and statistical significance: a practical guide for biologists. *Biol Rev Camb Philos Soc* 82: 591–605.