BJP British Journal of Pharmacology

DOI:10.1111/j.1476-5381.2011.01737.x
www.brjpharmacol.org

LETTER TO THE EDITOR

Statistical perspectives: all together NOT

Correspondence

Professor William G Hopkins, AUT University, Sport and Recreation, Akoranga Drive, Northcote, Private Bag 92006, Auckland 0627, New Zealand.
E-mail: whopkins@aut.ac.nz

This letter is being published in *The Journal of Physiology*, *Experimental Physiology*, the *British Journal of Pharmacology*, *Microcirculation*, and *Clinical and Experimental Pharmacology and Physiology*. It refers to a series of articles on best practice in statistical reporting.

An editorial under the title *Statistics: all together now, one step at a time* appeared in this and other physiological journals earlier this year (Drummond *et al.*, 2011) to announce a forthcoming series of 14 perspective articles on statistics. Two of the articles have now been published in six journals (Drummond and Vowler, 2011a; Drummond and Vowler, 2011b) and the third has started to appear (Drummond and Tom, 2011). The purpose of this letter is to identify what we consider to be substantial errors and omissions.

In the first article, *Show the data, don't conceal them* (Drummond and Vowler, 2011b), we agree with only the following two assertions. (For clarity, quoted assertions throughout this letter are shown in italics and are sometimes paraphrased for conciseness.)

- *Inspect dot plots of the original observations*. Yes, although such plots need not appear in the publication unless there is something unusual about the data.
- *'Dynamite-plunger plots' are inappropriate for repeated measurements*. Yes. The term *bar graph*, which exceeds *plunger plot* by a factor of 23 000 in a Google search, could have been introduced here. The term *line graph* is also worth noting for a plot in which means of repeated measurements are connected with line segments.

Next, several assertions we consider to be misleading.

- *Showing means and error bars would conceal the fact that a variable had a log–normal distribution*. This assertion may be true for the SEM but not the SD. A SD that in magnitude approaches or exceeds the mean is a good indication that the variable needs log transformation, as is a positive relationship between the SD and the mean when there are several groups (Altman and Bland, 1996). For this and other reasons, the error or error bar should always be a SD.
- *Error bars illustrating the 95% confidence interval (CI) should be included*. Use of CIs for effect statistics should be obligatory, but it is not appropriate to use CIs in dot or scatter plots or with means of original variables. The level of confidence need not be 95%. Sterne and Davey Smith (2001) suggested a default of 90%.
- *To define reference ranges a very large sample of normal values is needed*. The use of the word 'normal' here is confusing. A *random sample from the population* would be better. Also, depending on the distribution, a large or even modest sample size will suffice.

Now the assertions that are wrong.

- *Dynamite-plunger plots are never an appropriate way to plot the data*. We agree that bar graphs are inappropriate to display means of repeated measurements, but they are fine to convey magnitudes of differences in means in different groups, provided the error bars are SDs.
- *When data are skewed, the mean is not always the best choice of summary. An alternative and perhaps preferable index of scatter would be the 95% confidence limits . . . Then the 95% CI is chosen as the reference range*. Confidence limits or intervals are neither a measure of scatter nor a reference range. The authors presumably intended to refer to the use of quantiles. The usual quantiles for skewed data are the quartiles (lower quartile, median and upper quartile), often represented as box and whisker plots. The authors instead have suggested the 2.5th and 97.5th percentiles, which would be appropriate only to define the reference range for a clinical measure. There is then a confusing discussion about observations that fall outside the 95% reference range being 'abnormal'. It is unclear to us what readers are supposed to take from this section. Presumably the abnormality relates to flagging an individual's value in a clinical

setting; flagging of outliers in a research setting would require a reference range much wider than 95%.
- *The 95% confidence values for the mean indicate that we'd be 95% likely to get another estimate of the mean within this range, using another set of data samples from the same population.* Here we have the classic misconception about the meaning of the CI. In fact there is only an ~85% chance that another sample mean will fall within the 95% CI defined by the first sample (the exact value depends a little on the degrees of freedom). The 95% CI is a CI for the true (large-sample) mean, not the sample mean.

Finally, the omissions. These issues may be scheduled for future articles, but in our view they should have been at least mentioned in the first article.

- There is only one mention of logarithms, when data skewed to the right are defined parenthetically and inaccurately as *lognormal*. Why is there no mention of log transformation or use of logarithmic axes, and of the related issues of uniformity of effects and errors? The majority of the variables in physiological sciences and biomedicine generally need log transformation, and there are important implications for presentation of the data: with a log scale, with percentage or factor SDs and effects, and with back-transformed means of the log-transformed variable.
- Should the measure of variation shown with the mean be the SD or the SEM? The authors should have acknowledged this elephant in the room. The compelling evidence for exclusive use of the SD has been summarized elsewhere (Note 11 in Hopkins *et al.*, 2009).
- The focus of this first article is the distribution of the original values of the dependent variable, but the distribution of the residuals from the model used to derive effects, in particular the extent of non-uniformity, is more important.
- In any lesson on the presentation of data, one would expect to encounter the word *outlier* and pertinent advice on dealing with such data.

In the second article, *Data interpretation: using probability* (Drummond and Vowler, 2011a), our one major concern is the emphasis on null-hypothesis significance testing. It is unfortunate that an article devoted to probabilistic interpretation of effects lacks any mention of precision of estimation and confidence limits. Uncertainty is dealt with only as it relates to the null, and the key points reflect this emphasis. The authors have ignored the problems and discontent with this traditional approach to inference (e.g. Rozeboom, 1960; Cohen, 1994; Sterne and Davey Smith, 2001; Ioannidis, 2008; Ziliak and McCloskey, 2008; Stang *et al.*, 2010).

The example given for a virus killing a certain number of cells is belaboured and has unrealistic values. The calculation of probabilities is also inappropriate. The authors have assumed an outcome of eight cell deaths and have then calculated the probabilities of the various combinations of eight between the two cell types. But eight cell deaths is only one of many possible outcomes, so the probabilities are not correct for making inferences about the relative proportions of deaths. The correct analysis of these data – using a generalized linear model to compare the two proportions under the assumption of a binomial distribution – gives a narrower CI. The authors should have provided a more instructive example by computing and comparing confidence limits for a small sample size and a large sample size, using realistic data and taking into account what they did not mention, the smallest clinically or practically important difference.

The following statistical issues are also poorly articulated in the second article.

- *Estimate the probability that the observed data could have occurred by chance*. The chance of the observed data occurring is either 1 or 0, depending on your point of view. What the authors meant, presumably, is to estimate the traditional *P* value, which does not mean the probability that the observed data could have occurred by chance, even if the null hypothesis were true.
- *Statistical analysis allows the strength of this possibility [that our observations support a particular hypothesis] to be estimated. Because it's not completely certain, the converse of this likelihood shows the uncertainty that remains.* How should readers make use of this confusing statement? If we assume the traditional *P* value is implied as 'strength of possibility', a low value of *P* presumably implies in some sense a low possibility of the null hypothesis. Suppose $P = 0.14$. The 'converse' is presumably 0.86. So this is the uncertainty that remains? The uncertainty in what?
- *Consider the probabilities of more extreme data as well.* It does not make sense to speak of more extreme data, and in any case, more extreme than what? The null? The alternative hypothesis? The observed effect? It definitely makes sense to estimate the probabilities that the true value of the effect is greater than the smallest clinically or practically important positive and negative effect, but that is not what the authors intended here.
- *If you find 'no difference' this is no DETECTABLE difference.* It is incorrect to state 'if you find no difference', because to do so implies the classic misinterpretation of $P > 0.05$ as 'no difference'. To then say the difference should be interpreted as 'not detectable' still does not put things right, especially when the observed magnitude is substantial.
- *Here – as very often is the case – we should say 'at present, there could be an effect; the probability that there is no difference is not very small'.* It would have been a good idea to emphasize here that the *P* value is not the probability of no difference, because more than one of us got the impression from this article that the authors think it is. Also, the authors should have supplied better advice about what can be concluded with this kind of outcome. The probability that there is no difference is precisely zero: it is well known that with a big enough sample size, all effects are statistically significant; or to put it succinctly, the null hypothesis is always false. A conclusion along the following lines would be more informative: The uncertainty in the true effect allows for substantial positive and negative magnitudes; more data are required before we can infer a clear outcome.
- *To conclude 'this shows that there is no difference' here is to make perhaps one of the commonest errors in biology. A useful summary phrase is 'absence of evidence is NOT evidence of absence'.* While we agree about this endemic error, Altman and Bland's (1995) adage does not apply. When a researcher gets $P > 0.05$, is that absence of evidence? No, it

is actually some evidence of absence of the effect, in the sense of evidence for not rejecting the null hypothesis. Where most researchers or their statisticians get it wrong is indeed to then accept the null, when there is often good evidence for a substantial effect. Researchers can avoid this kind of mistake by estimating the uncertainty in the effect as confidence limits, then interpreting the magnitude of the confidence limits in relation to clinically, physiologically or practically important magnitudes, as illustrated at the end of the previous bullet point.

The third article (Drummond and Tom, 2011) is another exercise in rejecting the null hypothesis, this time in a controlled trial. Once again the opportunity to consider the uncertainty in an effect with respect to minimum important differences was lost: confidence limits were introduced only to determine whether they enclosed zero. In fact, the worked example would not lend itself easily to a consideration of the minimum important difference, which is defined here by a substantial change in chances of winning and is derived from the within-subject variability that elite frogs show from performance to performance (Hopkins *et al.*, 1999). A clinically relevant dependent variable, such as blood pressure or cholesterol, would have allowed the authors to introduce readers to the notion of a threshold change linked to a substantial change in morbidity or mortality.

There are various problems with the key points for the third article. They are almost all phrased in terms of null-hypothesis testing, with only one isolated point on CIs. Several points show imprecise or incorrect use of terms: *samples are often compared by first proposing that they could have come from the same population* (it is population parameters that are compared, not the samples, and the populations can be different: males and females, for example); *two random samples from the same population are unlikely to be the same* (are they ever the same?); *small samples are often imprecise* (how can a sample be imprecise?); *precision is also affected by the variability in a population* (variability of what, and which aspect of variability?).

The aim of the authors to provide clear non-technical guidance for authors on best practice in statistical testing and data presentation is commendable. However, the first three articles fall short of meeting this aim: the view of inference is flawed and outdated, the examples are inappropriate, there are serious errors and omissions, and the use of terms is imprecise. We hope the authors can address these deficiencies in future articles.

Will G Hopkins[1], Alan M Batterham[2], Franco M Impellizzeri[3], David B Pyne[4] and David S Rowlands[5]

[1]*AUT University, Auckland, New Zealand*, [2]*Teesside University, Middlesbrough, UK*, [3]*Schulthess Clinic, Zurich, Switzerland*, [4]*Australian Institute of Sport, Canberra, Australia, and* [5]*Massey University, Wellington, New Zealand*

References

Altman DG, Bland JM (1995). Statistics notes – Absence of evidence is not evidence of absence. BMJ 311: 485.

Altman DG, Bland JM (1996). Detecting skewness from summary information. BMJ 313: 1200.

Cohen J (1994). The earth is round (p <.05). Am Psychol 49: 997–1003.

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

Drummond GB, Vowler SL (2011a). Data interpretation: using probability. Br J Pharmacol 163: 887–890.

Drummond GB, Vowler SL (2011b). Show the data, don't conceal them. Br J Pharmacol 163: 208–210.

Drummond GB, Paterson DJ, McLoughlin P, McGrath JC (2011). Statistics: all together now, one step at a time. J Physiol 589: 1859.

Hopkins WG, Hawley JA, Burke LM (1999). Design and analysis of research on sport performance enhancement. Med Sci Sports Exerc 31: 472–485.

Hopkins WG, Marshall SW, Batterham AM, Hanin J (2009). Progressive statistics for studies in sports medicine and exercise science. Med Sci Sports Exerc 41: 3–13.

Ioannidis JP (2008). Why most discovered true associations are inflated. Epidemiology 19: 640–648.

Rozeboom WW (1960). The fallacy of the null-hypothesis significance test. Psychol Bull 57: 416–428.

Stang A, Poole C, Kuss O (2010). The ongoing tyranny of statistical significance testing in biomedical research. Eur J Epidemiol 25: 225–230.

Sterne JAC, Davey Smith G (2001). Sifting the evidence – what's wrong with significance tests? BMJ 322: 226–231.

Ziliak ST, McCloskey DN (2008). The Cult of Statistical Significance. University of Michigan Press: Ann Arbor.