npg

COMMENTARY

Statistics in Pharmacology

D Spina

Kings College London, Pharmaceutical Science Research Division, The Sackler Institute of Pulmonary Pharmacology, School of Biomedical and Health Science, Guy's Campus, London, UK

Statistics is an important tool in pharmacological research that is used to summarize (descriptive statistics) experimental data in terms of central tendency (mean or median) and variance (standard deviation, standard error of the mean, confidence interval or range) but more importantly it enables us to conduct hypothesis testing. This is of particular importance when attempting to determine whether the pharmacological effect of one drug is superior to another which clearly has implications for drug development and getting that next paper published in a respectable journal! Therefore, it is essential for pharmacologists to have an understanding of the uses and abuses of statistics. With this in mind, the *British Journal of Pharmacology* has commissioned a number of review articles to highlight the uses of statistics in experimental design and analysis.
British Journal of Pharmacology (2007) **152**, 291–293; doi:10.1038/sj.bjp.0707371; published online 9 July 2007

Pharmacology routinely employs statistics to help summarize data and, more importantly, to test hypotheses. This is a relatively simple matter when one is only interested in testing the Null hypothesis that two sample means are equal (H_0; $\mu_1 = \mu_2$). However, this type of experimental design and hence analysis does have a number of limitations. For example, one never simply investigates the effect of one dose of drug *in vivo*, and furthermore, it would not be sensible to have a control group for every drug dose group. Financial considerations notwithstanding, it is unethical to use large numbers of animals if a more appropriate experimental design can be implemented. Indeed, one can potentially reduce the number of animals used in an experiment by employing a better experimental design that, while a little more complex (for instance using factorial designs) has the advantage of answering more than one scientific question.

If you are confused about the most appropriate statistics for your experiment, you could simply talk to a statistician. However, it is clear that statisticians and pharmacologists often speak a different language. The statistician deals with uncertainties and calculates the probability of a particular event simply occurring by chance, and that is before you try and grasp the mathematics. In general, pharmacologists like to deal with certainties, for example, acetylcholine causes contraction of smooth muscle. Statisticians will test the Null hypothesis that sample means are equal, on the other hand,

a pharmacologist's starting premise is that drug X will increase blood pressure, so in effect, we want to test the Null hypothesis that sample means will be different. While these examples are frivolous, it is clear that we need to have a grasp of statistics to allow us to determine whether drug X is better than drug Y, for increasing blood pressure.

Unfortunately, the use of a subjective test like the 'obvious test' is not acceptable to journals and, therefore, we need to calculate the probability that the difference between drugs X and Y, the treatment effect, is greater than experimental error, and hence, the need for statistics.

Better experimental design is one potential benefit from learning the language of statistics. A stark reminder comes from an analysis of the relevance of investigating drug treatments in animal models as a precursor to investigational studies in man (Perel *et al.*, 2007). The conclusion drawn from this study was that animal experiments were biased or did not correctly model human disease and therefore were of little utility. However, there are some misconceptions concerning animal models and what they tell us about human disease. The majority of scientists do not set out to undertake clinical trials in animals *per se*. Rather, we test the hypothesis that drug treatment may or may not alter a biological response, which mimics some feature of the disease process we are studying. The so-called 'clinical trials' in animals are for all intent and purposes, proof of concept studies, which might give one confidence to proceed with the particular drug target of interest in man. Nevertheless, what this article and others (Festing, 2003) reveal is that experimental design could be vastly improved in order to reduce bias and the probability of obtaining a false positive or making misleading claims concerning the effectiveness of a particular treatment (that is, a type I error) in proof of concept studies. Therefore, it is vitally important that we

Correspondence: Dr D Spina, Kings College London, Pharmaceutical Science Research Division, The Sackler Institute of Pulmonary Pharmacology, 5th Floor Hodgkin Building, School of Biomedical and Health Science, Guy's Campus, London SE1 1UL, UK.
E-mail: domenico.spina@kcl.ac.uk

consider better experimental designs and analysis. This should not be difficult because we can adopt methods currently employed by our clinical colleagues who publish trial data. We should be adopting blinding strategies and randomization techniques when designing experiments and, of course, stating this fact in the Methods section of our MS. Furthermore, we should consider the use of factorial designs for many of our experiments for several very good reasons. Firstly, we are constrained to reduce the number of animals we use in experiments on ethical grounds. Hence, a factorial design is one such strategy that can be adopted to effectively reduce the total size of an experiment. For example, if we intend to examine the effect of different drugs or drug doses on a particular response *in vivo*, we only need a vehicle control group, not a control group for each drug or dose of drug. Factorial designs and hence analysis of variance will increase the power to detect differences between means (that is, to reject the Null hypothesis that treatment means are similar) because the degrees of freedom for the pooled error Sum of Squares term, will be greater and the error variance term relatively smaller, than the values obtained by application of a two sample Student's *t*-test between control and treatment groups (Wallenstein *et al.*, 1980; Festing, 2003). Third, the location of the difference between treatment means following a significant F ratio can then be addressed with either a planned comparison test (Armstrong *et al.*, 2000) or the more widely used *post hoc* analysis (Wallenstein *et al.*, 1980; Armstrong *et al.*, 2000), which enables the experimenter to ask several questions in one experiment. There are numerous *post hoc* tests available and all take into account the probability of making a type I error (false rejection of the Null hypothesis). Type I errors are usually made when multiple Student's *t*-tests are employed without making any correction for the number of Null hypotheses being tested. In fact, such data sets should be analysed using analysis of variance followed by the appropriate *post hoc* test (Wallenstein *et al.*, 1980).

Bearing this in mind, it is therefore timely that the *British Journal of Pharmacology* has commissioned a series of articles, which give invaluable advice for pharmacologists regarding uses and abuses of statistics, the first of which appears in this issue of the journal. In the first two articles of this series, Lew (2007a, b) provides important examples, which demonstrate how presentation of data in certain guises can be potentially misleading, the importance of using data transformation to normalize variance across treatment groups and how to employ analysis of variance for data analysis. An assumption of the analysis of variance is that the variance terms are homogenous, but often data sets collected in an experiment may give rise to non-homogenous variance across treatment groups, which invalidates an assumption of these parametric tests and can give rise to type I errors. Therefore, transformation of data is often usefully employed in this context and there are a number of different types of transformation readily available (Wallenstein *et al.*, 1980). Of course, there will be examples where data sets do not conform to a Gaussian distribution and it is not always possible to transform data. In these cases, non-parametric tests need to be utilized and will be described in later articles.

Lew (2007b) also gives an example of how good experimental design can be used to give the experimenter the best possible chances of rejecting the Null hypothesis. In this example, the advantages of using 'repeated measures' over single factor analysis of variance are discussed. This type of analysis lends itself to experimental designs that involve the repeated measurement within the same 'experimental unit' (that is, person, rat or cells from the same individual). The advantage of such designs is their greater power to reject the Null hypothesis and avoiding type II errors (failure to reject a Null hypothesis). A major source of experimental error in this type of design is the variability between 'experimental units' (such as genetic differences between subjects), which is removed from the calculation of the error variance term and, consequently, the chances of obtaining a significant F ratio is improved (Lew, 2007b). There are some assumptions for this type of analysis, which are not always met and therefore there is a tendency for type I errors to arise, but there are ways of getting around this problem with more conservative tests (Wallenstein *et al.*, 1980). Alternatively, depending on the nature of the serial measurement (time, concentration), one could calculate a summary statistic of the data in each individual, for example, the area under the time or concentration curve, EC_{50}, or peak response, and then use an appropriate statistical test to compare whether treatment means are different (Matthews *et al.*, 1990). Unfortunately, one often sees pairwise comparisons between dose or concentration levels between a control and treated group. This is highly erroneous and must be avoided because the comparisons are not always independent and therefore invalidate one of the assumptions of the Student's *t*-test, which states that each comparison must be independent. For example, if the means of three treatment groups were A, B and C, then the following two statements $A = B$ and $B > C$ can be described as two independent comparisons and it must therefore follow that $A > C$; however, this comparison is not independent. In this case, it would be far better to obtain a summary statistic of the serial measurement, as described above, and then compare the treatment means with an appropriate parametric test (Matthews *et al.*, 1990).

Statistics can be very helpful in formulating experimental design and drawing appropriate inferences from the collected data. If we employ better design and analysis, we will reduce the risk of making misleading claims and provide greater confidence that our proof of concept studies may translate into man. This of course assumes that we have chosen the correct drug target in the first instance and this will only be known from human clinical trials. While statistics can tell us whether we should accept or reject the Null hypothesis, the overriding question that we must ask ourselves, in the case of rejecting the Null hypothesis, is whether the magnitude of this difference has any biological relevance? For this, we need our pharmacological hat.

References

Armstrong RA, Slade SV, Eperjesi F (2000). An introduction to analysis of variance (ANOVA) with special reference to data from clinical experiments in optometry. *Ophthalmic Physiol Opt* **20**: 235–241.

Festing MF (2003). Principles: the need for better experimental design. *Trends Pharmacol Sci* **24**: 341–345.

Lew MJ (2007a). Good statistical practice in pharmacology: problem 1. *Br J Pharmacol* **152**: 295–298 (this issue).

Lew MJ (2007b). Good statistical practice in pharmacology: problem 2. *Br J Pharmacol* **152**: 299–304 (this issue).

Matthews JN, Altman DG, Campbell MJ, Royston P (1990). Analysis of serial measurements in medical research. *BMJ* **300**: 230–235.

Perel P, Roberts I, Sena E, Wheble P, Briscoe C, Sandercock P *et al.* (2007). Comparison of treatment effects between animal experiments and clinical trials: systematic review. *BMJ* **334**: 197–200.

Wallenstein S, Zucker CL, Fleiss JL (1980). Some statistical methods useful in circulation research. *Circ Res* **47**: 1–9.