British Journal of Pharmacology (2008) 155, 797–803

0007–1188/08 $32.00
www.brjpharmacol.org

REVIEW

On contemporaneous controls, unlikely outcomes, boxes and replacing the 'Student': Good statistical practice in pharmacology, problem 3

MJ Lew

Department of Pharmacology, University of Melbourne, Parkville, Victoria, Australia

This paper is intended to assist pharmacologists to make the most of statistical analysis and avoid common errors. A scenario, in which an experimenter performed an experiment in two separate stages, combined the control groups for analysis and found some surprising results, is presented. The consequences of combined controls are discussed, appropriate display and analysis of the data are described, and an analysis of the likelihood of erroneous conclusions is made. Comparisons between data from separately conducted experimental series are hazardous when there is any possibility that the properties of the experimental units have changed between the series. Experiments that have been performed independently should be analyzed independently. Unlikely or surprising results should be treated with caution and a high standard of evidence should be required, and verification by repeated experiments should be performed and reported. Box and whisker plots contain more information than plots more commonly used to display for qualitative variables and should be used where the sample size is large enough (say, $n \geqslant 5$). In most biomedical experiments the observations are not random samples from large populations as assumed by conventional parametric analyses such as Student's *t*-test, and so permutation tests, which do not lose their validity when a sampled population is non-normal or when the data are not random samples, should frequently be used instead of Student's *t*-tests.
British Journal of Pharmacology (2008) **155**, 797–803; doi:10.1038/bjp.2008.350; published online 22 September 2008

Keywords: permutations tests; box and whisker plots; Bayesian logic; unlikely outcomes; historical controls; combined controls

The problem

A pharmacologist involved in cardiovascular research was asked to test the potential of a novel antioxidant compound to reduce cardiac cell damage after myocardial infarction. The routine assay involved occlusion of a coronary artery of anaesthetized rats for a controlled period, followed by reperfusion. The rats are killed some time later and the necrotic portion of the heart is determined by staining.

Rats were randomly allocated to control (vehicle injection) or test (5 mg kg^{-1} of the test compound during reperfusion), and 10 observations in each group were obtained before data analysis with Student's *t*-test. The compound appeared to exacerbate the cardiac damage, and the effect was statistically significant ($P = 0.037$).

The pharmacologist was then asked to test the compound at a lower dose, and so the experiment was repeated with a test dose of 2.5 mg kg^{-1}. This time the compound caused an unexpected but statistically significant reduction in cardiac damage ($P = 0.019$).

Correspondence: Dr MJ Lew, Department of Pharmacology, University of Melbourne, Parkville, 3010, Victoria, Australia.
E-mail: michaell@unimelb.edu.au
Received 26 June 2008; revised 7 August 2008; accepted 13 August 2008; published online 22 September 2008

It was concluded that the drug had a bidirectional action, protective at low doses and deleterious at high doses. The data are shown in Figure 1.

Questions:

(1) What danger signs can be discerned from the description and in the data as displayed?
(2) How might the display of the data be made more informative?
(3) The conclusion is unsafe. Why?
(4) What statistical analyses should be used for these experiments?

Analysis of the problem

What danger signs can be discerned from the description and in the data as displayed?

A clear danger sign in the problem as presented is that the study is described as having been conducted in two stages, but the data figure implies only one stage. Another clue that

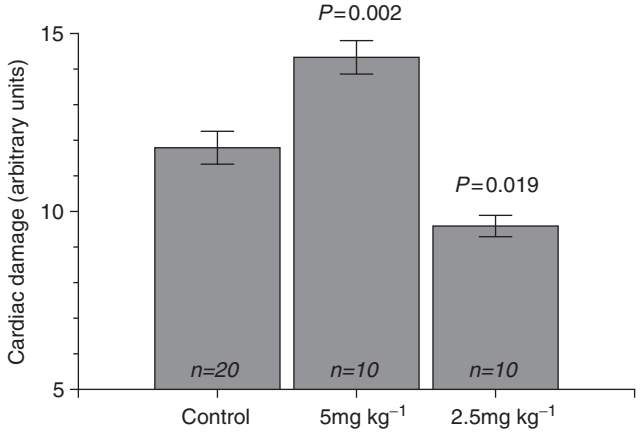

Figure 1 Results of the experiments described in this problem. The indicated *P*-values are for comparisons between the test groups and control. Error bars are s.e.

something is amiss in the experiment is the inconsistent *P*-value for the first test that can be seen by comparing the textual description and the data figure. The textual description is deliberately sufficient to give an accurate impression of the sequence of events, but its narrative style is unusual for a modern methods section and it may be chastening to consider whether the same impression would be gained if it was written in a more typical manner. Without the narrative description 'A pharmacologist… was asked to… was then asked to...' the only clues available to the reader would be that the control group has twice the number of observations but the same s.e. as the test groups. The unusual arrangement of the data in the figure, which has the larger dose to the left of the smaller, may also be taken as a danger sign but, frankly, so little thought appears to be given to designing the graphical display of data in many basic research papers that such a feature may mean nothing at all.

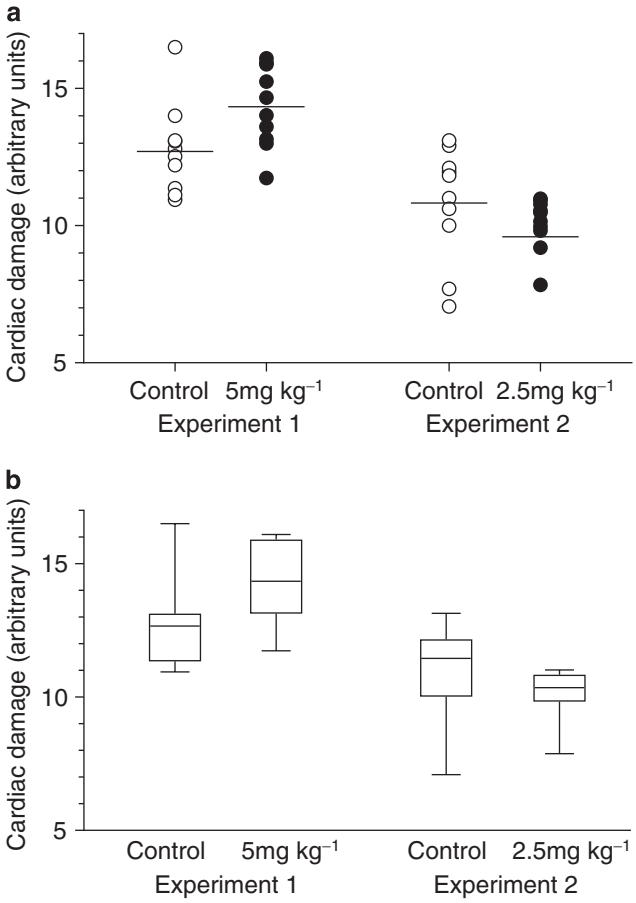

Figure 2 Results of the experiments of this problem replotted to make the experimental design and the outcomes clear. (**a**) graph showing individual data points ($n = 10$ for each treatment) and their means. (**b**) the same data shown as a box and whisker plot. In this plot the whiskers extend to the largest and smallest data points, the box extends from the upper quartile to the lower quartile and is crossed by a line at the median of the data.

How might the display of the data be made more informative?
The data are displayed as bars, a widely used graph type, but one that displays relatively little information. Other important information is obscured by the arrangement of the graph, which implies that three treatments were tested in a single experiment and, arguably, by the fact that the bars do not start at zero.

The issue of whether bar graphs should have zero baselines is somewhat controversial. One argument is that although bar graphs present information regarding the magnitude of means via the *y*-axis scale, they more powerfully show the relative magnitudes of the various means via the relative sizes of the bars—their 'visual mass' or bulk. Having a non-zero baseline distorts the relative bulks of the bars and may mislead the viewer. A contrary school of thought, expressed, for example, by Good and Hardin (2003), argues that the differences between, rather than the absolute magnitudes of, the means are what is important in a bar graph and therefore chopping off the bottom of the bars is a useful way of emphasizing the point of interest. I find the first argument more compelling. In cases where the *relative* magnitude of the means is unimportant, the data should be displayed more effectively with a plot where the means are represented with a line or a dot rather than the top of a bar.

Another feature of bar graphs is worth pointing out here: bar graphs treat data *x*-values as qualitative categories, thereby degrading quantitative *x*-values. In the case of the current data, the *x*-values are quantitative—0, 2.5 and 5 mg kg^{-1} doses—but because each experiment had only two doses, the degradation of the *x*-axis to qualitative categories does not significantly reduce the information. However, there are many examples where it would; for example, one does not have to look very hard to find whole concentration–response curves displayed using bar graphs (in other journals, of course).

Figure 2 shows two better alternatives for displaying the data, each showing explicitly the two stages of experimentation. The first panel shows all of the data points individually, along with each group's mean. That graph is similar to the improved display in the previous paper in this series (Lew, 2007), but in this study, the larger group sizes lead to some overlap of the points. Because of the overlap, using a box and whisker plot to present a 5-point summary of the data may

be a better option. In a box and whisker plot the box extends from the upper quartile to the lower quartile of the data (it is sometimes called the interquartile range) to enclose the central half of the data. A line drawn across the box indicates the median of the data and whiskers extend from the box to the largest and smallest values (with optional end caps). As the box and whisker plot summarizes the data into 5 points, it is not an appropriate plot when a data set is very small. When the data consisted of exactly 5 points the summary values may be taken as the data points themselves. Thus, for small sample sizes a simple plot of the data points along with their means is a better option. However, where the data set is as large as in this problem ($n = 10$) the box and whisker plot provides an uncluttered and informative display and deserves to be used more often.

The failure of pharmacological researchers to universally adopt the box and whisker plot may be the result of several interrelated obstacles. First, there are many different variants of the box and whisker plot—their originator, John Tukey, presented two versions in his original description (Tukey, 1977) and shortly thereafter described several more variants (McGill *et al.*, 1978). Researchers may pass them over as a display option for fear of using the wrong variant, and the multiplicity of variants combined with their rarity in the pharmacological literature means that explicit explanations are needed to accommodate many readers. Another problem—a surprising one—is that there are many different approaches to the calculation of the upper and lower quartile values that inconsistently yield different values (Langford, 2006). Different graphing programs sometimes give different looking box and whisker plots from the same data. I have tested ProFit (Quantum Soft), Prism (Graphpad) and Excel (Microsoft) but found that they do not consistently agree on quartiles. Each provides 'correct' quartiles insofar as they divide the data into quarters, but the quartiles of discontinuous distributions may be non-unique and each program appears to use a different algorithm for their calculation. The magnitude of discrepancies between quartiles is small when the samples are large and so the use of different algorithms may not be a serious problem, but ideally all graphing programs would at least provide the option of using quartiles that are calculated in the manner described by Tukey (1977) when constructing box and whisker plots.

The conclusion is unsafe. Why?

In the improved display of the data (Figure 2) it can be immediately seen that there were two separately controlled experiments conducted. It is also clear that any effect of the higher dose of the test compound is marginal compared with the variability in the data, and that no effect could be confidently ascribed to the lower dose. Importantly, there is little evidence for the most interesting part of the conclusion, namely that the drug had a bidirectional effect.

Comparison of Figure 2 shows that the original conclusion is unsafe because it was based on a likely misapprehension about the results that were obtained. The drug probably had no effect at all at the lower dose, and the effect at the higher dose was likely of questionable biological significance, a 12% increase, despite being significant in a statistical sense ($P = 0.037$). As the putative effect of the high dose was in an unexpected direction, the statistical result should be treated with a degree of caution. The previous sentence will sound reasonable to some readers, but its sentiment may not be familiar to all and therefore it deserves expansion. The first expansion offered is in plain language—well, as plain as I can manage anyway—and the second will be more exact (and might act as a very basic primer on the logic of Bayesian statistics). Either explanation alone should suffice and therefore feel free to skip the second, but, with a bit of luck, the first will help make the second accessible.

Expansion 1. Consider the situation where, in a long series of independent experiments, one experiment returns a significant outcome. There are two possible explanations: there might be a true effect or the result might be a false positive. Which of those explanations is more likely? Assuming that the experiments are well powered, most of the experiments in which there is a true effect will return a significant result, and, assuming for convenience the use of $P < 0.05$ as the critical threshold, only 1 in 20 of the experiments in which there is no effect should yield a significant result (which would be a false positive). The balance between the false-positive outcomes and the true positives therefore depends on not just the critical threshold of P selected by the experimenter, but also on the relative proportion of experiments in which there is a true effect to observe. If most of the experiments in the long series that we are considering have a real effect then very few of the significant outcomes will be false positives; false positives are only possible when there is no effect. If, on the other hand, there is a true effect in only a few of the experiments, then the false positives will be a substantial portion of all of the significant results, even to the extent that the long-run probability of a significant result being a false positive can be much higher than the 0.05 critical value. Thus, the likelihood that any particular significant result is a false positive is high when a true effect is unlikely.

The paragraph above describes a very artificial circumstance in which the experiment in question is a member of a large series of experiments with a known, or approximately known proportion of treatments having a true effect. In conventional pharmacological experiments, we would frequently be loath to specify in advance the proportion of our experiments that involve a true treatment effect. However, there are some circumstances in which we might hazard a guess. If an experimenter already has preliminary observations that support the possibility of an effect then the same effect would not be considered unlikely in a subsequent experiment. Conversely, in a large random screening study it is common for a substantial proportion of the apparent positive outcomes ('hits') to turn out to be false positives making it important for hits to be validated with additional experimentation. The narrative description of the experiments in this problem make it sound like they may have been the first test of the drug in question on myocardial damage, and therefore there may not have been any relevant pilot studies. However, it may be assumed that information from other assays led the experimenter to expect a protective

effect, and therefore the study is not quite like a random screening exercise. However, if a result can be described as 'unexpected', as it is in the problem, then it is reasonable to treat it is as having been unlikely.

Expansion 2. This version is based on the same notional series of independent experiments, but treats the concepts more formally and quantitatively. Thus, I will specify the experimental design power as well as the false-positive error rate, that is, the notional experiments are all conducted with the power to detect a true effect 80% of the time (a false-negative error rate, β, of 20%), and a false-positive error rate, α, of 0.05. When the *P*-value from any of the experiments is less than 0.05, the result is considered to be evidence against the null hypothesis. All experiments are conducted in circumstances in which the null hypothesis, H_0, is either true or false, and so possible outcomes of such experiments can be illustrated with a probability tree diagram (Figures 3a and b). It can be seen that the probabilities on the terminal branches of the tree diagram come directly from the experimental design. For example, the probability of obtaining evidence against a true H_0 is the probability of making a false-positive error, α. Similarly, the probability of failing to obtain evidence against a false H_0 is the false-negative error probability, β, and the power of the experiment to detect a true effect is shown as the probability of finding evidence against a false H_0, 1-β. Tree a shows the relevant probabilities symbolically, and tree b substitutes the values.

Now, let us use the tree diagram to explore the situations described in the previous explanation, that is, where a real

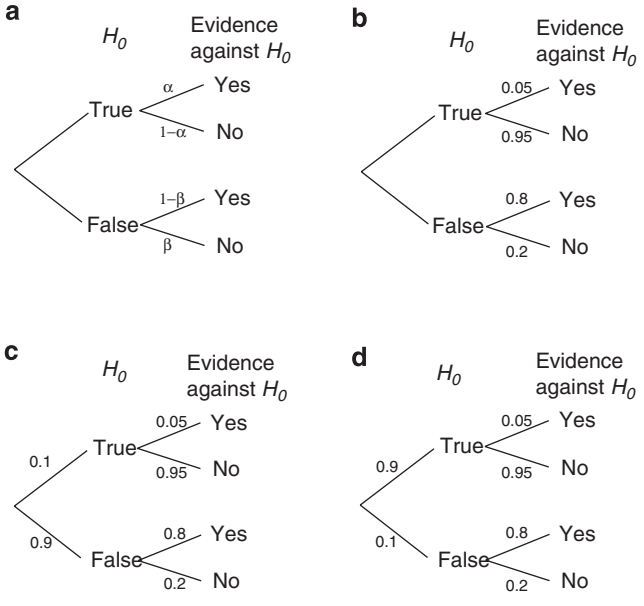

Figure 3 Tree diagrams that illustrate the probabilities of obtaining the four possible combinations of null hypothesis (H_0) condition (true or false) and experimental outcome (sufficient or insufficient evidence against H_0). (**a**) generic tree with conventional symbols for the probability of a false-positive outcome (α) and a false-negative outcome (β). (**b**) tree with the probabilities of false-positive and false-negative outcomes as specified in the text. (**c**) tree for circumstances in which a true effect (false H_0) is common (9 times out of 10). (**d**) tree for circumstances in which a true effect (false H_0) is rare (1 time out of 10).

effect is common as well as rare. To do so we need to assign probabilities to 'common' and to 'rare' real effects. Those probabilities do not need to be extreme for the outcomes to be interesting, therefore assume that common corresponds to a probability of 9 out of 10, and rare is 1 out of 10 (Figures 3 c and d). Those tree diagrams can now be used to illustrate the calculation of the probabilities of obtaining reliable and unreliable evidence against H_0. In all cases, the probability of obtaining evidence against H_0 is the sum of the probabilities of getting to 'Yes' in the trees. In the case in which a true effect is common, it can be calculated as $(0.1 \times 0.05) + (0.9 \times 0.8) = 0.725$. That value is dominated by the probability of a correct outcome, $(0.9 \times 0.8) = 0.72$, and the probability of a false-positive outcome is very low, $(0.1 \times 0.05) = 0.005$, and therefore the probability that a significant outcome represents a false positive is 0.005/0.725, less than 1%. However, in the case in which a true effect is rare, the probability of obtaining evidence against $H0$ is less, $(0.9 \times 0.05) + (0.1 \times 0.8) = 0.125$, and the probability of a false-positive outcome makes up about a third of the total probability. In that case, the probability that a significant outcome represents a false positive is 0.045/0.125, or 36%. Clearly, it is risky to make a very strong claim regarding an unexpected significant effect!

Readers who are confused at this point should reflect on whether they hold the mistaken idea that the *P*-value from a statistical test is the probability that the result is a false positive. In fact, it is the probability of obtaining data at least as extreme as those observed when the null hypothesis is true. The *P*-value does reflect the probability that any given result is a false positive, but that probability is also dependent on other factors as described above.

The aforegoing explanation of why caution should be used in the interpretation of surprising results uses Bayesian logic, but is not couched in the mathematical terms that are usually used in textbook chapters about Bayesian statistics. A very brief formal description may be useful for some readers—others can skip this bit. Answering the question of whether a significant result is more likely to be a correct result or a false positive requires calculation of the probability that the null hypothesis is true in light of the experimental evidence against it. The famous Bayes' theorem provides a way to calculate $P(\mathrm{A}|\mathrm{B})$, the probability of one stochastic event, A, given another, B:

$$P(\mathrm{A}|\mathrm{B}) = \frac{P(\mathrm{B}|\mathrm{A})P(\mathrm{A})}{P(\mathrm{B})}$$

What we want is very similar, $P(H_0|E)$, which can be read as the probability of the null hypothesis, H_0, (being true) given that there is evidence against it (event *E*). That can be had by a simple substitution of terms into the formula above to yield

$$P(H_0|E) = \frac{P(H_0|E)P(H_0)}{P(E)}$$

The probability of contrary evidence (occurring) given H_0 is the conventionally obtained *P*-value, or the chosen cutoff value for significance, 0.05 in this case. The probability of the null hypothesis, $P(H_0)$, is called the prior probability, and

it is the value that was arbitrarily set to 0.1 in Figure 3c, and 0.9 in Figure 3d. The denominator, $P(E)$, is the probability of obtaining the evidence independent of the truth of the null hypothesis. $P(E)$ is calculated as $P(E|H_0).P(H_0)+P(E|H_A).P(H_A)$, where H_A is the alternative hypothesis or the negation of the null hypothesis. Thus, substituting the numerical values, where the probability of the null hypothesis being true was 0.1 or 0.9, respectively, we get:

$$P(H_0|E)=\frac{0.05\times0.1}{0.05\times0.1+0.8\times0.9}=0.007$$

or

$$P(H_0|E)=\frac{0.05\times0.9}{0.05\times0.9-0.8\times0.1}=0.36$$

Those probabilities are exactly as we calculated before, and therefore it can be seen that the Bayesian approach is not as unnatural as might be assumed on the basis of the gulf between those who describe themselves as 'Bayesians' and those who do not. For the problem presented here it really does not matter whether a formal Bayesian approach is followed, or the informal approach used in the first explanation. Either way the conclusion is the same: a significant effect in an unexpected direction does not allow firm conclusions to be made. If the result is interesting then the experiment should be repeated.

What statistical analyses should be used for these experiments?
The first experiment was intended to answer the question 'does the drug affect the amount of damage caused by cardiac ischaemia and reperfusion?' but the aim of the second experiment is less clear. It might have been an attempt to validate the surprising result of the first experiment by repetition, although in that case the use of a different dose was unfortunate. However, it might have been an attempt to find out whether the lower dose had an effect similar to the higher dose. In that case, the use of a dose only 50% lower was unfortunate because one should expect similar doses to give similar outcomes. Perhaps the second experiment was intended to serve both purposes. In any case, the experiments were designed and run as two separate series, and cannot be reliably analyzed as if they were one.

A two-sided Student's *t*-test is almost universally used by pharmacologists for comparisons of means in experiments such as those in this problem. However, the question about 'what analyses to perform' should not be answered completely on the basis of convention, as it may not give the best solution, and there are a couple of issues to consider first. We will start with the consideration of whether a one-sided test might be used. The fact that the drug was expected to reduce cardiac damage from ischaemia and reperfusion means that the original hypothesis was at least arguably one-sided, and some might take that as sufficient justification for using a one-sided (one-tailed) test and thereby to gain twice the power to detect a real effect. However, there are disadvantages to using one-sided tests as well. When a one-sided test is used, any apparent effect of the intervention in the opposite direction to that hypothesized (such as the one that may have occurred in this case) has to be totally ignored. One cannot decide to use a one-sided test and then after seeing the data alter the direction of the effect, or even change to a two-sided test because the long-term false-positive error rate would be increased. That leads to another disadvantage of one-sided tests: they are sometimes regarded with a great deal of suspicion by referees, and it may be difficult to publish inferences that are supported by one-sided tests. The only circumstances in which one-sided tests can be recommended is where effects are physically or biologically possible only in one direction, or where an effect in one direction would be of no interest whatsoever, no matter how juicy. Neither of those circumstances is common in pharmacological experimentation, and certainly neither is true in this particular case. Thus, any test used in these experiments should be two-sided.

Now, we turn to consideration of what specific tests would be appropriate to test for differences between the treatments in the problem. On the basis of convention, Student's *t*-test is an obvious choice. However, the underlying premise of Student's *t*-test, and other conventional parametric tests for differences between means, is that the data are randomly sampled from large, normally distributed populations with equal variance. Populations? What populations? In my experience, animals available to a researcher come in small batches from a central animal holding or breeding facility, are allocated in a manner that is largely outside of the experimenter's control, and may include both littermates and unrelated animals. Rather than being randomly sampled from large populations treated and untreated, the animals are a single sample of convenience from a population that may not be much larger than the sample itself. Those conditions are potentially disastrous to the logic of parametric statistical analysis. Fortunately, we can get around the discrepancy between theory and reality by assuming that the convenience sample is itself an effectively random subset of a much larger population of potential animals—perhaps real animals that might potentially be allocated to the experimenter, or animals that might potentially exist and be allocated. That assumption is probably reasonable much of the time. At least it is from the point of view of utility and acceptability, as can be gauged by the widespread and accepted use of Student's *t*-test and the like for analysis of experiments on convenience samples. However, the assumption is not always safe and its limitations become evident, for example, when the notional population is repeatedly sampled over time. Inevitably the biological properties of experimental animals change from time to time as a consequence of many factors: the animals may have seasonal biological adaptations; foodstuffs may vary seasonally; the animal house may change its routines or food suppliers; the animals may become diseased or, worse, may be treated pharmacologically for disease without the investigator's knowledge. Readers can no doubt think of other issues that would mean that a sample drawn today would be different from a sample drawn 2 months ago not only for whole animals, but for tissue samples and cultured cells as well. Therefore, even if it is reasonable to pretend that convenience samples are equivalent to random samples from a large population, it is not reasonable to assume that they are

samples from an invariant population. This issue invalidates the use of 'historical controls' in experiments in which the subjects are more biological than coins or dice and means that contemporaneous control and test experiments are essential.

The discrepancy between the two control groups in the experimental data presented here can easily be explained as a change in the population properties during the interval between the first and second experimental series. The assumption that the sample of convenience approximates a random sample may be reasonable *within* each series of experiments, but it is not reasonable for both of the series *together*. Therefore, a Student's *t*-test for the effect of the 5 mg dose compared to the first set of control data would be acceptable, and so would a Student's *t*-test of the second series, but the analysis done in the problem that combined the control groups is altogether unsound.

Even though Student's *t*-tests can be made to serve for analysis of these data, there is a better approach that does not require any assumption of sampling from a large, normally distributed population. The permutations test is a randomization test for differences between group means which makes no assumptions regarding random sampling, and can be used whenever a Student's *t*-test might be used (Colquhoun, 1971; Ludbrook and Dudley, 1998). It is based on a model in which the data are assumed to be randomly allocated (that is, randomized) to treatment groups rather than randomly sampled from large populations. In effect, permutations tests use the observed data as the entire population and thus there is no need to make assumptions about sampling from notional or real populations. The standard approach to experiments in laboratory-based pharmacology resembles random allocation much more closely than it resembles random sampling—the convenience sample is randomly split into different treatment groups, but not randomly sampled from larger populations—and thus permutations tests are more natural analyses for these data than are Student's *t*-tests. Permutations tests ask the question 'how unusual is the arrangement (permutation) of the values within the data compared with all possible arrangements (permutations)?' and provide a probability result, which is simply the number of possible data arrangements that produce a difference between group means as large or larger than that observed, divided by the total possible number of arrangements. (Interestingly, that probability is of exactly the form of the classical definition of probability that we have from Laplace, who said that probability is the number of favourable outcomes divided by the total number of outcomes.) The null hypothesis tested by the permutations test is that the arrangement of the data into their groups is random, as would be the case if they were randomized into treatment groups and if the treatment(s) were ineffective. That is different to the null hypothesis of the Student's *t*-test, which is that the populations from which the data were sampled have the same means, distribution and variance. Although direct inference from the parametric test extends to populations, direct inference from the permutations test reflects only the values in the samples tested. However, it is easy to generalize the latter inference by inductive argument, and there is arguably some difficulty associated with the generalization of inference about populations that are almost always ill defined and often only notional. Permutations tests have power equivalent to Student's *t*-test when the sample size is more than five, even when the data are normally distributed and ideal for the parametric test. When the data are not ideal for the parametric test the results from the permutations test will inevitably be more reliable. Thus, there are good reasons for preferring the permutations test over the Student's *t*-test (Colquhoun, 1971; Edgington, 1995; Ludbrook and Dudley, 1994, 1998; Good, 2005).

The calculation of a permutations test is conceptually straightforward—simply examine all possible arrangements of the data—but, as is often the case, the practicalities of the calculation can be tedious. The total number of possible permutations quickly gets very large as the sample size is increased. For example, with two groups of $n=10$ observations the number of permutations is 184 756, and with $n=12$ it is 2 704 156. Happily, computers can make quick work of the calculations, and several different software packages will do the job (including one for Windows and Macintosh written by the author, available at http://www.pharmacology.unimelb.edu.au/Statboss/Site/Home.html). Testing the results of the two experimental series independently for an effect of the drug with a permutations test yielded $P=0.039$ for the 5 mg dose, and $P=0.30$ for the 2.5 mg dose. Those values are similar to what would be obtained from Student's *t*-test, 0.037 and 0.30, respectively (note that the correct value for the 2.5 mg dose is that given here, the value in the problem text is erroneous because it used the combined, $n=20$, control group). The similarity of the results from the two test types is to be expected because the conventional Student's *t*-test is valid in circumstances in which the sampling was non-random only insofar as it provides an estimate of the *P*-value that would be obtained from a permutations test (Edgington, 1995 pages 10–13; Fisher, 1936).

Conclusions and recommendations

(1) Comparisons between data from separately conducted experimental series are hazardous when there is any possibility of changes in the properties of the experimental units between the series. Experiments that have been performed independently should be analyzed independently.
(2) Unlikely or surprising results should be treated as unlikely and a high standard of evidence should be required. Verification by repeated experiments should often be performed and reported.
(3) Box and whisker plots contain more information than plots more commonly used to display for qualitative variables and should be used when the sample size is large enough (say, $n \geqslant 5$).
(4) In most biomedical experiments the observations are not random samples from large populations as assumed by conventional parametric analyses such as Student's *t*-test. Permutations tests do not lose their validity when a sampled population is non-normal or when the data are not random samples, and should frequently be used instead of Student's *t*-tests.

Conflict of interest

The author states no conflict of interest.

References

Colquhoun D (1971). *Lectures on biostatistics*. Clarendon Press: Oxford.

Edgington ES (1995). *Randomization tests*, 3rd edn. Marcel Dekker Inc.: New York.

Fisher RA (1936). 'The coefficient of racial likeness' and the Future of Craniometry. *Journal of the Royal Anthropological Institute* **66**: 57–63 (freely available online from the R.A. Fisher digital archive: http://digital.library.adelaide.edu.au/coll/special/fisher/papers.html).

Good PI (2005). *Permutation, Parametric, and Bootstrap Tests of Hypotheses*. 3rd edn. Springer Verlag: New York.

Good PI, Hardin JW (2003). *Common Errors in Statistics (and How to Avoid Them)*. Wiley: Hoboken, NJ.

Langford E (2006). Quartiles in elementary statistics. *Journal of Statistics Education 14*, www.amstat.org/publications/jse/v14n3/langford.html.

Lew MJ (2007). Good statistical practice in pharmacology, problem 2. *Br J Pharmacol* **152**: 299–303.

Ludbrook J, Dudley H (1994). Issues in biomedical statistics: statistical inference. *Aust NZ J Surg* **64**: 630–636.

Ludbrook J, Dudley H (1998). Why permutation tests are superior to *t*- and *F*-tests in biomedical research. *Am Stat* **52**: 127–132.

McGill R, Tukey JW, Larsen WA (1978). Variations of box plots. *Am Stat* **32**: 12–16.

Tukey JW (1977). *Exploratory Data Analysis*. Addison-Wesley: Reading, MA.