British Journal of Pharmacology (2007) 152, 299–303
© 2007 Nature Publishing Group All rights reserved 0007–1188/07 $30.00
www.brjpharmacol.org

REVIEW

Good statistical practice in pharmacology Problem 2

M Lew

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

Background and purpose: This paper is intended to assist pharmacologists to make the most of statistical analysis and in avoid common errors.
Approach: A scenario is presented where an experimenter performed an experiment to test the effects of two drugs on cultured cells. Analysis of the results, expressed as percentage of control, by a one-way ANOVA yielded $P = 0.058$ and the experimenter concluded that neither drug was effective. The data were expressed as percentage of control because of pairing of the data within each experimental run, a common feature in cell culture experiments. Such data can be analysed with potentially more powerful ANOVA methods equivalent to the paired *t*-test. Monte Carlo simulations are presented to compare the power of relevant analyses.
Results: For data correlated within experimental run (i.e. paired values), transformation to percentage of control improved the power of a one-way ANOVA to detect a real effect, but a randomized block ANOVA (equivalent to a 2-way ANOVA with experiment and treatment as factors) using the raw values was substantially more powerful. The randomized block ANOVA performed well even with uncorrelated data, being only marginally less powerful than the one-way ANOVA.
Conclusions and implications: A randomized block ANOVA is far superior to the one-way ANOVA with correlated data, and with uncorrelated data it is only marginally less powerful. Thus where there is, or might reasonably be, such a correlation (e.g. relatedness among the data within a single experimental run, or within a multi-well culture plate, or within an animal, et cetera), use the more powerful randomized block ANOVA rather than one-way ANOVA.
British Journal of Pharmacology (2007) **152,** 299–303; doi:10.1038/sj.bjp.0707372; published online 9 July 2007

Keywords: data interpretation; statistical analysis; data transformation; research design; Monte Carlo method; analysis of variance; randomized block

Problem

A pharmacologist performed an experiment to determine whether cultured cells respond to two drugs. The experiment was conducted using a stable cell line plated onto Petri dishes, with each experimental run involving assay of responses in three Petri dishes: one treated with drug 1, one with drug 2 and one serving as control. The experiment was run six times on successive days. As each run of the experiment consisted of all treatments, the pharmacologist decided to express the data as a percentage of the control value to minimize the contribution of day-to-day variation to the noise in the data.

Summary results subjectively indicated the effectiveness of drug 1 compared to control (Figure 1), but a conventional one-way analysis of variance (ANOVA) did not allow the experimenter to confidently reject the null hypothesis that the three groups are the same, with $P = 0.058$. It was concluded that neither drug was effective in this experiment.

Question: Are there any errors in the experimental analysis or the interpretation of the results?

Analysis of problem 2: Getting the power you paid for

Data display
The first step in data analysis is inspection of the data. The data were provided in the problem only in a summarized form – means and error bars – and the meaning of the error bars was not explicitly mentioned, but as most readers probably assumed, they are standard errors of the mean. The data were 'normalized' to be a percentage of the control value from each experiment, and thus the relative variability in the different treatments is distorted and the actual values are obscured. That presentation of data is common, but it hides more than it shows. We can do much better. There are few enough data points, so one could simply provide the

Correspondence: Dr M Lew, Department of Pharmacology, University of Melbourne, The University of Melbourne, Parkville, Vic 3010, Australia.
E-mail: michaell@unimelb.edu.au
Received 20 December 2006; accepted 30 May 2007; published online 9 July 2007

Table 1 Raw data obtained from the experiment

	Control	*Drug 1*	*Drug 2*
Experiment 1	1147	1169	1009
Experiment 2	1273	1323	1260
Experiment 3	1216	1276	1143
Experiment 4	1046	1240	1099
Experiment 5	1108	1432	1385
Experiment 6	1265	1562	1164

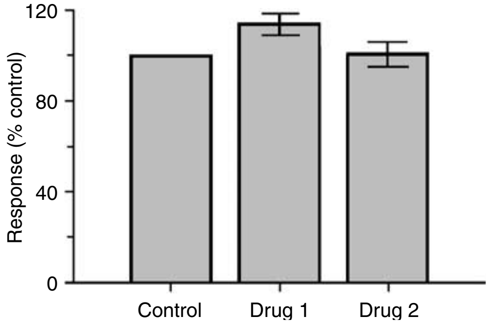

Figure 1 The effects of drug 1 and drug 2 on the responses of cultured cells as described in the text.

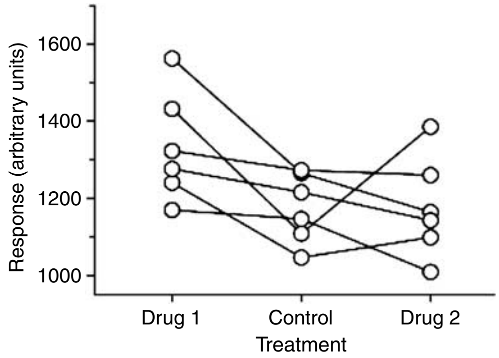

Figure 2 An information-rich display of the data in Table 1. Lines link data points that were obtained from each experimental run. Compare this figure with Figure 1.

values numerically (Table 1), but a well-designed graphical display is usually helpful. Figure 2 shows all of the values with the data points from within an experiment connected to make their relationship clear. The treatments are arrayed spatially with the control in the centre, so that the slope of the lines properly reflect the direction of any potential effect. It can be seen that in every case, the drug 1 values are larger than their corresponding control value but only two of the six experiments showed a larger value for drug 2 treatment than control. There is far more information in Figure 2 than there is in Figure 1, and yet it takes exactly the same space.

Statistical analysis

The original analysis of these data used a parametric one-way ANOVA, which is a frequent choice for analysis of data like these. That ANOVA assumes that the values are independent and normally distributed and that there is an even distribution of variability in the groups. We cannot be certain that the normal distribution assumption is valid because tests for normality need far more data points to be useful. Nonetheless, inspection of the raw data (Figure 2, not Figure 1) would suggest no major problem with that assumption, particularly given that ANOVA is not very badly affected by minor deviations from the normal distribution. Unfortunately, the data as originally analysed certainly do clash with the assumption of equal variability across the treatments – the control group had zero variability – and so that version of the data should not be analysed with a parametric ANOVA. The raw data groups do seem to have equal variability and so we could base our analysis on the raw data. However, the reason that the experimenter chose to express the data as percent of control in the first place was the desire to 'minimize the contribution of day-to-day variation to the noise in the data' – in other words, to make the analysis more powerful. Are we going to lose analytical power by forgoing that transformation in order to meet the criteria for the parametric ANOVA? Not at all. In fact, we will end up with a more powerful analysis.

The percent of control transformation uses the relatedness of the values obtained within each experimental run to minimize the overall variability but unavoidably compromises any parametric analysis that included the control group. The relatedness of the values can be exploited without invalidating the parametric analysis by using an ANOVA that is equivalent to a paired *t*-test (related sample *t*-test). Such an ANOVA is a two-way ANOVA with experiment and treatment as the two factors, or a randomized block design ANOVA, and is sometimes called a repeated measures ANOVA. The nomenclature is confusing, but the calculations, and thus the results, are the same in each case. The experimental design used in these experiments is a randomized block with each experimental run being a block. The name 'block' comes, like many other descriptors in ANOVA, from agricultural experiments where blocks of land are divided into subplots and each subplot is randomly assigned one of the treatments (for example, fertiliser or seed type). In our experiments, the subplots (Petri dishes within a day) were each subjected to a different drug treatment, which we assume was randomly allocated. Thus we have six blocks each with three subplots. An ANOVA for randomized blocks is therefore appropriate, and it consists of an ANOVA where the variation between the blocks is segregated from the total variation before the calculation of the treatment effect. Exactly the same calculation is carried out for a two-way ANOVA where the factors of interest are drug treatment and experimental day. The data we are dealing with are repeated measures in that they are multiple measurements from a single experimental unit (in this case, they are the measurements from a single run of the experiment), although that is not the most common usage of the term repeated measures. (Note: there are two distinct forms of repeated measures – measurements that are 'repeated' in parallel like those in this case, and measurements that are repeated serially or sequentially over time. ANOVA for the serial repeated measures can be differentiated from that described in this article, and generally involves the Green-

Table 2 Randomized block analysis of variance table for the data shown in Table 1

Source	*d.f.*	*Sum of squares*	*Mean squares*	*F*	P*-value*
Between columns (treatments)	2	99 122	49 561	5.27	0.027
Between rows (blocks)	5	134 190	26 838	2.85	0.074
Residual	10	94 024	9402		
Total	17	327 336			

house–Geisser or Huynh–Feldt corrections for unequal correlations). Some computer programs for ANOVA will require you to call the analysis one thing, and others another, but they should give equivalent results. I will call the analysis a randomized block ANOVA, because it refers to the particular experimental design as well as the analysis.

The randomized block ANOVA table obtained from the data in Table 1 is shown in Table 2 (analyse the data in Table 1 yourself to see whether your preferred statistics program does the same analysis under a different name). It can be seen with this analysis that there has been a significant effect of the treatment – $P = 0.027$ for between columns – an outcome that is a ‘success’ where the one-way ANOVA gave a ‘failure’ or, at best, an unconvincing result of $P = 0.058$. It looks like the randomized block analysis has given the better outcome, but we cannot really be sure on the basis of this one experiment which type of ANOVA gave the *right* result. If there were no real drug effects then the one-way ANOVA gave the right result and the randomized block ANOVA resulted in a false claim of an effect (a type I error). Conversely, if there was really an effect then the one-way ANOVA resulted in failure to claim a true effect (a type II error), whereas the randomized block ANOVA gave the right answer. It seems reasonable to prefer the randomized block analysis, because it is designed specifically for data like these, but is it really much better than the one-way ANOVA? To find out, we have to do some simulations.

If you wish to simply accept without evidence my conclusions and recommendations then you may skip the Simulations section, but if you interested in seeing some empirical evidence, or in the possible magnitude of any differences in ANOVA performance, then read on.

Simulations

In the discussion above, it was noted that it is not possible to say which type of ANOVA supported the correct conclusion from the experiments, because it is not known whether the treatments were truly effective. To be confident that the ‘significant’ outcome from the randomized block ANOVA is better than the ‘nonsignificant’ outcome from the one-way ANOVA, we have to see how well the two types of analysis perform when we do know the real answer. Monte Carlo simulations allow us to do just that. We can find out how often the analyses give the wrong result when there is no real drug effect, and how often they give the wrong result when there is an effect – the false-positive (type I) and false-negative (type II) error rates, respectively. The power of an analysis to detect an effect is one minus the false-negative error rate. We want as few type II errors as possible, but at the same time we do not want to make too many false claims of success and so we also want good control of the type I errors.

For the Monte Carlo simulations, 10 000 sets of normally distributed pseudo-data were generated with the same arrangement as in Table 1, three columns of six values, and each set analysed by the randomized block and one-way ANOVAs. To mimic the experimental data, which are expected to be correlated within rows because of blocking in the experimental design, the pseudo-data sets were made to be correlated within rows, with an average correlation coefficient of 0.5. In the first set of simulations, the data were made to have no true difference between the column means and so those simulations provide us with the false-positive error rates. A second set of simulations was generated with the mean of one column being one standard deviation (s.d. of the population, usually denoted as σ) greater than the others—in other words, with a true effect—to determine the false-negative error rates. The error rates for one-way ANOVA were also determined under those conditions, but with the pseudo-data sets expressed as a percentage of their respective control values.

The results from the simulations (Table 3) show that the false-positive error rate of the one-way ANOVA was lower than that of the randomized block ANOVA. That might be thought of as a positive trait, but it indicates that the one-way ANOVA sets the bar for ‘significance’ higher than needed for the nominal type I error rate, $\alpha = 0.05$. Statisticians would describe the test as behaving in an overly conservative manner with these data, and there is a corresponding elevation of the false-negative error rate to 0.72 (that is, the power to detect the real effect was only 0.28, about one out of four). One-way ANOVA performed better when the data were expressed as a percentage of the control value, with a false-negative error rate of 0.52, and so it seems that the strategy of the experimenter to improve his or her analysis by reducing the inter-experiment variability was at least partially successful. However, the randomized block ANOVA behaved even better, with the false-positive error rate exactly matching the nominal $\alpha = 0.05$ and a false-negative error rate of 0.42. Those results show that the randomized block design ANOVA correctly detected a one s.d. difference between means 58% of the time compared to the one-way ANOVA result of only 28% for the raw data and 48% for the percent control data. The randomized block ANOVA is clearly superior to the one-way ANOVA when the data are correlated.

The results in Table 3 indicate that we should strongly prefer the randomized block ANOVA for analysis of data with correlated values within the rows, but setting up an experiment with a randomized block design does not guarantee that the data within a block will be correlated to

Table 3 Results from simulations to determine the performance of one-way and randomized block ANOVAs with data correlated within rows

	Type I errors *(α = 0.05)* *(no true effect)* *frequency*	*Type II errors* *(α = 0.05)* *(true effect = 1 × σ)* *frequency*
One-way ANOVA	0.01	0.72
One-way ANOVA (% control)	0.04	0.52
Randomized block ANOVA	0.05	0.42

Abbreviation: ANOVA, analysis of variance.
Type I error frequency was determined from the number of times that $P<0.05$ was returned by the ANOVA in 10 000 simulations where there was no true effect, and type II error frequency from the number of times the ANOVA returned $P>0.05$ when there was a real effect equal to the population standard deviation, σ. Ideal values would be type I error rate equal to the cutoff P-value, α, and type II errors as low as possible.

Table 4 Results from simulations to determine the performance of one-way and randomized block ANOVAs with independent data

	Type I errors *(α = 0.05)* *(no true effect)* *frequency*	*Type II errors* *(α = 0.05)* *(true effect = 1 × σ)* *frequency*
One-way ANOVA	0.05	0.65
One-way ANOVA (% control)	0.04	0.74
Randomized block ANOVA	0.05	0.68

Abbreviation: ANOVA, analysis of variance.
Type I error frequency was determined from the number of times that $P<0.05$ was returned by the ANOVA in 10 000 simulations where there was no true effect, and type II error frequency from the number of times the ANOVA returned $P>0.05$ when there was a real effect equal to the population standard deviation. Ideal values would be type I error rate equal to the cutoff P-value, α, and type II errors as low as possible.

an important degree. Thus, we need to know whether there is a high price to pay for applying the randomized block ANOVA to uncorrelated data. To find out, simulations were run as before, but with independent, uncorrelated values in the rows. The results (Table 4) show that for those pseudo-data sets, the one-way ANOVA on the raw data behaved best – predictably, as those data are exactly what are assumed by that analysis – but the randomized block design was very nearly as good. The one-way ANOVA on the data expressed as percent of control was substantially less powerful than either of the other analyses. Thus, whereas there is a large advantage in using a randomized block ANOVA instead of a one-way ANOVA where the data are correlated, there is only a very minor disadvantage in doing so when the data are entirely uncorrelated. It is reasonable, therefore, to apply the randomized block ANOVA whenever the experimental design is of that form, even if the expected correlation among the values is slight. Expressing the data as a percent of control improves the power of a one-way ANOVA when there is a correlation among the values within an experiment, but reduces power when there is no correlation, and it never performs as well as the randomized block design ANOVA.

There is an issue raised by the ANOVA results in Table 2 that is worthy of some brief consideration. As the analysis is identical to a two-way ANOVA, the table include two P-values, 0.027 for the treatment effect, and 0.074 for the block effect. The former is less than the conventional, but arbitrary, cutoff of 0.05, and so we call it 'significant', but the latter is not. One might think that because the between blocks comparison did not yield a significant P-value there would be no advantage in using the randomized block ANOVA instead of a one-way ANOVA, but that is definitely not the case. The between blocks variability (sum of squares = 134 190) is a substantial fraction of the total variability (sum of squares = 327 336), hardly 'insignificant'. On theoretical grounds, it can be quite definitively said that the $P>0.05$ for the between rows comparison does not mean that the blocking was ineffective. For a blocked experimental design, our default expectation has to be that the data will be correlated. The null hypothesis for the between rows comparison in Table 2 is therefore that the data are correlated and any comparison between rows is effectively a 'reverse experiment' that needs to be tested using equivalence tests (Lew, 2006). The effect of this might be made clear by consideration of the results of the simulations used to generate Tables 3 and 4. In the simulations above, the P-value for the between blocks comparison was a relatively unreliable indicator of whether there was a real correlation between the values within a block. Exactly 5% of the randomized block ANOVAs gave $P<0.05$ for the between rows comparison when the data were independent: a perfect agreement between empirical results and the nominal type I error rate for that comparison. However, when there was a real correlation within the rows (Table 3 data), the between rows comparison yielded $P<0.05$ only 54% of the time. If we used the between rows comparison to decide whether the randomized block ANOVA should be used, we would be wrong 46% of the time. Thus, if the experiment is designed as randomized blocks then you should use a randomized block ANOVA whether the effect of blocking looks 'significant' in the ANOVA table or not.

Interpretation of the results

Now that the effectiveness of the relevant ANOVA designs has been established, we can turn our attention back to the original experiment. On the basis of the one-way ANOVA, the experimenter concluded that 'neither drug was effective in this experiment'. However, that conclusion was based on the results of an analysis that we know to be inappropriate in two ways: the data did not meet the assumptions built into the analysis; and the analysis did not make full use of the power inherent in the experimental design. In any case, a conclusion that neither drug was effective would not follow from a result of $P>0.05$, even if the analysis had been perfect. A P-value greater than our predetermined cutoff for significance means that the results were not unusual enough to support a conclusion that any differences between the group means were not due to chance alone. Being unable to conclude that there is a difference is not the same as concluding that there is no difference. If one or both drugs did have an effect, it was too small relative to the noise to be discriminated with the desired level of confidence with the one-way ANOVA.

Of course, we should be paying attention to the results of the randomized block ANOVA instead of the one-way ANOVA. That analysis gave a result for the between treatments comparison of $P = 0.027$, low enough for us to conclude with reasonable confidence that one of the treatment groups was different to at least one of the others. However, while that ANOVA result tells us that something interesting has happened, it does not say which group is different. Inspection of the data suggests that the mean of the drug 1-treated group is higher than the other groups. To find out, we need to do some more statistics to make specific comparisons that are commonly called *post hoc* tests. Those tests are not the topic of this article, so we will simply say that Dunnett's test is generally a good way to compare one specified column to all others. With a control group and two different drugs, Dunnett's test allows us both of the meaningful comparisons (drug 1 to control and drug 2 to control) without wasting power on the pointless comparison of drug 1 to drug 2. One could use Bonferroni's correction for multiple *t*-tests, but Dunnett's test gives good control of type I errors with more power because Bonferroni's correction is conservative. Dunnett's test gives $P = 0.033$ for drug 1 and $P > 0.99$ for drug 2. Thus, our conclusion is that drug 1 was effective, but any possible effect of drug 2 was far too small relative to the noise to be discriminated.

It is worth noting that the effect of drug 1 was not very large. The effect might be important, but it could equally well be scientifically insignificant despite being statistically significant. Importance cannot be decided by the statistical analysis. If the experimenter was interested in discriminating drug effects as small as that observed, then we should probably say that the experiment was under-powered – the simulations with correlated data (Table 3) showed that the randomized block ANOVA was able to detect a true effect as large as the s.d. only 58% of the time when there are six values in each group. In other words, this experiment might have had the power to detect a difference of that size only about every second time, even when the analysis makes full use of the power of a randomized block experimental design.

Recommendations

The results of this investigation support the following conclusions, assuming that the underlying 'raw' data are from normally distributed populations.

(1) Do not express the data as percent of control before analysis using ANOVA. That transformation makes any data set violate the ANOVA assumption of equivalent variances and can reduce the power of the analysis. It may be desirable to show the data as percent control, say for clarity or consistency, but there is no requirement that the data be analysed in the form that is displayed.

(2) Where there is, or might reasonably be, a within row correlation between the values (that is a correlation among the data within a single experimental run, or within a multi-well culture plate, or within an animal, et cetera), choose the more powerful randomized block ANOVA rather than one-way ANOVA. Of course, as always, choose the type of analysis before seeing the results of any analysis!

(3) Ignore the *P*-value for the between blocks comparison in the randomized block ANOVA table.

Reference

Lew MJ (2006). When there should be no difference – how to fail to reject the null hypothesis. *Trends Pharm Sci* **27**: 274–278.