



FORUM

Data pooling and type I errors: a comment on Leger & Didrichsons

STEPHEN H. JENKINS

Department of Biology, University of Nevada, Reno

*(Received 20 November 2000; initial acceptance 12 March 2001;
final acceptance 19 June 2001; MS. number: AF-13)*

Leger & Didrichsons (1994) discussed the implications of data pooling for descriptive and inferential statistics in animal behaviour. They defined data pooling as 'an analytic procedure in which multiple samples of an individual's behaviour are treated as independent events' (Leger & Didrichsons 1994, page 823). Machlis et al. (1985) had suggested that this practice is common in animal behaviour despite its violation of the assumption that samples used to calculate descriptive statistics for groups or to compare two or more groups should be independent. In contrast to Machlis et al. (1985), Leger & Didrichsons (1994) claimed that data pooling is appropriate when within-subject variance is greater than between-subject variance or sample sizes are equal across subjects, and may be advantageous when it is difficult to study large numbers of individuals.

Leger & Didrichsons' (1994) paper has been frequently cited as a justification for data pooling in studies of animal behaviour (a search using Science Citation Index yielded 63 citations as of 23 May 2001). However, their analysis was flawed in three important ways. (1) They downplayed the fact that variation within and among individuals may be inherently interesting. (2) They ignored standard statistical methods for dealing with studies in which multiple observations are collected from more than one individual. (3) Their claim that pooling does not increase the probability of type I error in comparisons of two or more groups is incorrect.

There is increasing interest in the causes and consequences of individual variation in behaviour and other traits of organisms (Clark & Ehlinger 1987; Hayes & Jenkins 1997). One important reason for this interest is that individual variation is the raw material on which natural selection acts, so that studying individual variation may be a powerful tool for learning more about the process of microevolution. Studies of individual variation may become as important for evolutionary ecology in the future as studies of geographical variation have been in

Correspondence: S. H. Jenkins, Department of Biology/314, University of Nevada, Reno, NV 89557, U.S.A. (email: jenkins@unr.edu).

the past (Foster & Endler 1999). Leger & Didrichsons (1994) took the traditional approach of viewing variation within and among individuals as a problem to be overcome in describing and comparing behaviour of groups, rather than recognizing that such variation may be an interesting subject of study in itself for what it can reveal about opportunities for natural selection within populations.

Leger & Didrichsons (1994) discussed several case studies involving simulated and real data in which there were multiple observations for more than one subject in each of one or more groups. As they clearly stated, the observations for a single subject are not independent. This means that such data should be analysed with a nested analysis of variance (ANOVA) if they are normally distributed (Underwood 1997), or with hierarchical generalized linear modelling (GLM) if not (Bryk & Raudenbush 1992; Osborne 2001). The latter is a modern, flexible approach for dealing with mixed-effect models with normal or non-normal data in a nested structure. Failing to treat the data with nested ANOVA or hierarchical GLM leads to the classical problem of pseudoreplication (Hurlbert 1984). If two or more groups are being compared, then use of nonindependent observations within groups leads to inflation of degrees of freedom for statistical tests and increased probability of falsely rejecting the null hypothesis (type I error). The multiple observations for a single individual essentially represent subsamples; use of such subsampling can yield a more precise estimate of mean values for each individual, but the individual means should be used for comparisons among groups. Besides leading to appropriate statistical tests of differences among groups, nested ANOVA and hierarchical GLM provide a basis for estimating components of variance among and within individuals, which is an important foundation for studying individual variation (Lessels & Boag 1987; Boake 1989; Hayes & Jenkins 1997).

Leger & Didrichsons (1994, page 828) recognized the problem of nonindependence in stating that 'pooling

does indeed increase the chance of rejecting the null hypothesis'. They went on to claim that 'the rapidly asymptotic nature of the F -distribution' implies that this increased probability of type I error makes little difference for large enough sample sizes (e.g. five observations of each of 20 subjects in each of two groups). However, an explicit analysis of one of their examples (Leger & Didrichsons 1994: Figure 9) vitiates this claim. In their Figure 9, Leger & Didrichsons (1994) presented a hypothetical data set with four observations for each of three subjects in each of two groups. There was a small difference in the mean values for the two groups and much greater variance within subjects than between subjects. I examined the probability of type I error for this situation by simulated sampling from normal distributions. In these simulations, the null hypothesis was true (i.e. the means for each group were set equal to zero). I first drew mean scores for each subject from a normal distribution with a mean of 0 and a variance of 0.1, representing low between-subject variance. I then drew multiple observations for each subject from a normal distribution with the mean determined in the previous step and a variance of 1.29, matching the high ratio of within-subject variance to between-subject variance used by Leger & Didrichsons (1994). I repeated this process 10 000 times, each time doing one-way ANOVAs to test for apparently significant differences between the two groups using pooled data (all individual observations) and aggregated data (subject means), as done by Leger & Didrichsons (1994). In these simulations, an apparently significant result at $\alpha=0.05$ implied a type I error, because group means were identical.

I ran these simulations for all combinations of 5, 10, or 20 subjects and 5, 10, or 20 observations per subject. For aggregated data, the highest probability of type I error was 0.0531, very close to the nominal level of 0.05. For pooled data, probabilities of type I error were about 0.08 for five observations per subject, 0.13 for 10 observations per subject, and 0.20 for 20 observations per subject, and were independent of the number of subjects (Fig. 1a). I repeated the simulations with between-subject variance exceeding within-subject variance, and found probabilities of type I error close to 0.05 for aggregated data but ranging from 0.4 to 0.7 for pooled data (Fig. 1b). Thus, regardless of the asymptotic properties of the F distribution and contrary to Leger & Didrichsons (1994), pooling may cause substantially elevated probabilities of falsely rejecting null hypotheses. This problem is more serious when between-subject variance is large compared to within-subject variance (Fig. 1b), but also exists when the opposite is true (Fig. 1a). The problem disappears only if between-subject variance equals zero (Barcikowski 1981).

In addition to these hypothetical data, Leger & Didrichsons (1994) discussed actual data on acoustic properties of crying by human infants at 1 and 6 months of age. The two ages differed significantly for 14 of 26 acoustic variables when pooled data were analysed and for eight variables when aggregated data were analysed. Analyses of pooled and aggregated data gave different results for eight variables. In seven cases, there was a significant difference with pooled data but not with

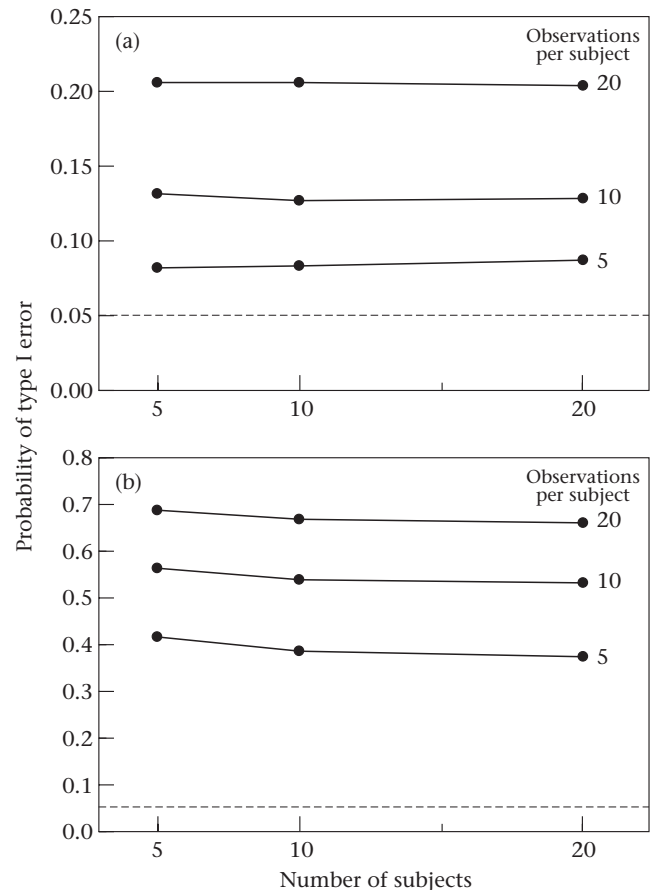


Figure 1. Probability of type I error using pooled data for comparing two groups with 5, 10 and 20 subjects and 5, 10 and 20 observations per subject. In both panels, means for the two groups were equal and simulated data were drawn from normal distributions for 10 000 trials as described in the text. In (a) within-subject variance was 1.29 and between-subject variance was 0.1; in (b) within-subject variance was 0.1 and between-subject variance was 1.29. The dotted line shows the nominal significance level of 0.05.

aggregated data; in one case the opposite was true. Leger & Didrichsons' (1994, page 829) interpretation of this was that 'aggregation may produce a nonconservative test of the null hypothesis', especially when within-subject variance exceeds between-subject variance. A nonconservative test of a null hypothesis is a test that produces a type I error probability greater than the specified significance level of the test (e.g. greater than 0.05). As demonstrated above, use of pooled data may lead to a nonconservative test while use of aggregated data does not. A more appropriate interpretation of the results of this study is that some of the apparently significant differences between ages simply represented type I errors, which are more likely with pooled data (14 potential cases) than with aggregated data (8 potential cases).

In summary, students of animal behaviour should not use Leger & Didrichsons (1994) as a guide for treating data when they have multiple observations for more than one subject in one or more groups. There is a standard statistical approach, nested ANOVA, which is appropriate for these kinds of data and provides interesting and

important information about relative magnitudes of within-subject and between-subject variance. For data that do not meet the normality assumption of ANOVA but have nested structure, the relatively new approach of hierarchical generalized linear modelling can be used for a wide range of situations (Osborne 2000). Pooling data, as recommended by Leger & Didrichsons (1994) for some situations, increases the probability of type I error, possibly to very high levels (Fig. 1).

I thank D. W. Leger and L. Thomas for valuable comments on the manuscript.

References

- Barcikowski, R. S. 1981. Statistical power with group mean as the unit of analysis. *Journal of Educational Statistics*, **6**, 267–285.
- Boake, C. R. B. 1989. Repeatability: its role in evolutionary studies of mating behavior. *Evolutionary Ecology*, **3**, 173–182.
- Bryk, A. S. & Raudenbush, S. W. 1992. *Hierarchical Linear Models: Applications and Data Analysis Methods*. Newbury Park, California: Sage Publications.
- Clark, A. B. & Ehlinger, T. J. 1987. Pattern and adaptation in individual behavioural differences. In: *Perspectives in Ethology*. Vol. 7: *Alternatives* (Ed. by P. P. G. Bateson & P. H. Klopfer), pp. 1–47. New York: Plenum.
- Foster, S. A. & Endler, J. A. (Eds) 1999. *Geographic Variation in Behavior: Perspectives on Evolutionary Mechanisms*. New York: Oxford University Press.
- Hayes, J. P. & Jenkins, S. H. 1997. Individual variation in mammals. *Journal of Mammalogy*, **78**, 274–293.
- Hurlbert, S. H. 1984. Pseudoreplication and the design of ecological field experiments. *Ecological Monographs*, **54**, 187–211.
- Leger, D. W. & Didrichsons, I. A. 1994. An assessment of data pooling and some alternatives. *Animal Behaviour*, **48**, 823–832.
- Lessels, C. M. & Boag, P. T. 1987. Unrepeatable repeatabilities: a common mistake. *Auk*, **104**, 116–121.
- Machlis, L., Dodd, P. W. & Fentress, J. C. 1985. The pooling fallacy: problems arising when individuals contribute more than one observation to the data set. *Zeitschrift für Tierpsychologie*, **68**, 201–214.
- Osborne, J. W. 2000. Advantages of hierarchical linear modeling. *Practical Assessment, Research & Evaluation*, **7**(1): <http://ericae.net/pare/getvn.asp?v=7&n=1>.
- Underwood, A. J. 1997. *Experiments in Ecology: Their Logical Design and Interpretation Using Analysis of Variance*. New York: Cambridge University Press.