An assessment of data pooling and some alternatives

DANIEL W. LEGER* & INTA A. DIDRICHSONS†
**Department of Psychology and School of Biological Sciences, University of Nebraska, Lincoln, NE 68588-0308, U.S.A.*
†The Gallup Organization, 301 S. 68th Street, Lincoln, NE 68510, U.S.A.

*(Received 7 July 1993; initial acceptance 10 July 1993;
final acceptance 13 September 1993; MS. number: A6519R)*

Abstract. Data pooling is an analytic procedure in which multiple samples of an individual's behaviour are treated as independent events. Although common in animal behaviour research, data pooling has been discredited because it may violate statistical assumptions. Four data sets were analysed in both their pooled and unpooled forms. Pooling did not bias results provided that either intra-subject variance exceeded between-subject variance or *N*s were equal. Between-groups tests of significance were affected in the same way as descriptive statistics, and as long as intra-subject variance exceeded between-subject variance, pooling did not increase the probability of a type I error. Utmost care must be taken to sample individuals from populations and behaviour from those individuals in an unbiased manner.

Behavioural research depends on drawing representative samples from two populations. First, and most commonly recognized, one must obtain an unbiased sample of individuals from the population of interest. The sample must include individuals of all appropriate demographic classes and they must not be more conspicuous or otherwise more readily available than the population from which they are drawn.

The second sort of sample comes from the population of all possible behavioural expressions that can be performed by each individual. For instance, in a study on song rate in birds, a 10-min data collection period might be deemed appropriate, but a large number of potential data collection periods are available for each individual. The sample of behaviour can be biased by obtaining data only at certain times of day, in certain locations, etc., unless one is only concerned with those more limited realms. Sampling of individuals and sampling of their behaviour are independent processes and one may obtain unbiased samples of one, both, or neither.

Assuming that one's research protocol results in unbiased samples of individuals and behaviour, interesting methodological questions can be asked. Can more than one behavioural sample be taken from an individual? If so, can they be treated in the same manner as one might treat single samples from each of one's subjects? Or should multiple behavioural samples from an individual be aggregated into a mean score prior to further analysis? Under what circumstances, if any, can multiple behavioural samples from an individual be treated as being equivalent to single samples from multiple individuals? These are some of the questions addressed in this article.

When an individual's behaviour is measured more than once and such measurements are treated in the same way as measurements obtained from different individuals, the 'pooling fallacy' can potentially occur (Machlis et al. 1985). Multiple samples from individuals may not be independent, thus violating statistical assumptions. These extra data points also increase degrees of freedom for the error term, thus tending to increase the chance of falsely rejecting the null hypothesis.

Machlis et al. (1985), in a widely cited, influential paper, claim that pooling is common in studies of animal behaviour, implying that much of our knowledge of behaviour is suspect. Similar concerns have been voiced by Hoekstra & Jansen (1986), Kroodsma (1989, 1990) and Beal & Khamis (1990). Appropriately, animal behaviourists have become leery of data pooling.

If data pooling is a questionable procedure, why have researchers used it? Pooling is especially common in field studies of endangered species or small populations because small populations often require multiple samples from each individual to achieve an adequate sample size. Pooling also is

0003–3472/94/100823+10 $08.00/0

used when only a limited subset of the members of a population can be observed, such as when the researcher is restricted to a blind or when extensive habituation to the observer's presence is required. In addition, pooling also may occur in large, unmarked populations if one's sampling procedure results in some individuals being measured more than once. Given that such constraints are unlikely to disappear, knowledge of the effects of pooling would be valuable so that we can begin to assess how our conclusions are influenced.

The primary question about pooling is whether the behaviour population can be represented in an unbiased manner by sampling *N* individuals once each, or by sampling, for instance, half as many individuals twice each. We would have the same total number of data points in each case, but their origins would differ. The answer to this question seems to depend on the variability between subjects and the variability within subjects over time. For instance, in a hypothetical population of individuals that all have the same mean and variance on the behaviour of interest, it would make no difference whether one obtained 100 data points by sampling 100 individuals one time each, 50 individuals twice, or, at the other extreme, one individual 100 times. We find the latter case extremely worrisome, however, because we do not deal with populations of individuals with the same mean and variance. But if such were the case, both the single- and multiple-sample procedures would in fact provide an unbiased sample of the population.

In contrast to the preceding hypothetical case, consider the opposite extreme: a population of individuals whose mean scores differ greatly from one another, but who have extremely small intra-individual variance. The more disparate the individuals' mean scores are from one another and the more discontinuous their score ranges, the greater the necessity of sampling a larger proportion of the population, but because of small intra-individual variability, there would be little value in collecting more than one behavioural sample from each subject.

Of course, real populations may not approach either extreme. But consideration of these extreme possibilities leads to some interesting hypotheses regarding choices among sampling strategies. When intra-subject variance is large relative to between-subject variance, then sampling a limited number of individuals more than once may provide an unbiased estimate of the population mean and variance. When intra-subject variance is small relative to between-subject variance, sampling a large proportion of individuals from the population may be necessary to obtain an unbiased estimate of population parameters.

Procedures for avoiding pooling may introduce problems of their own. Measuring each individual only once may limit sample sizes to the point that the sample is not representative of the population. Aggregation (i.e. representing each individual by its mean score) may not be the answer either, especially if the distribution of the behaviour is bimodal, in which case the mean score for an individual will be one that rarely occurs. Aggregation can also give the erroneous impression that unexplained variance is smaller than it actually is.

In the following studies we test the assertion that pooling is a reliable procedure provided that intra-subject variance exceeds between-subject variance. Furthermore, we show that procedures other than pooling may not be reliable under certain circumstances.

STUDY 1: A SIMULATION

Methods

We constructed two populations consisting of 10 individuals, each of which was sampled 10 times. Both populations had means of 100, but their intra- and between-subject variances differed. These can be seen in Fig. 1. In Population 1, the intra-subject variance was 1·9 times greater than the between-subject variance. In Population 2, the intra-subject variance was only 0·3 times as great as the between-subject variance.

We used five methods to compute estimates of the means and standard deviations of the two populations. (1) Complete pooling: we used all 100 scores. (2) Limited pooling: we randomly selected 10 scores, allowing no more than two scores from the same individual. In 10 replications of this procedure, the number of individuals with two samples ranged from two to four. Therefore, not all 10 subjects were represented in any one replication. (3) Single sampling: we randomly selected 10 scores, one from each individual. This procedure was replicated 10 times. (4) Limited aggregation: we randomly selected scores such that each individual was represented by at least

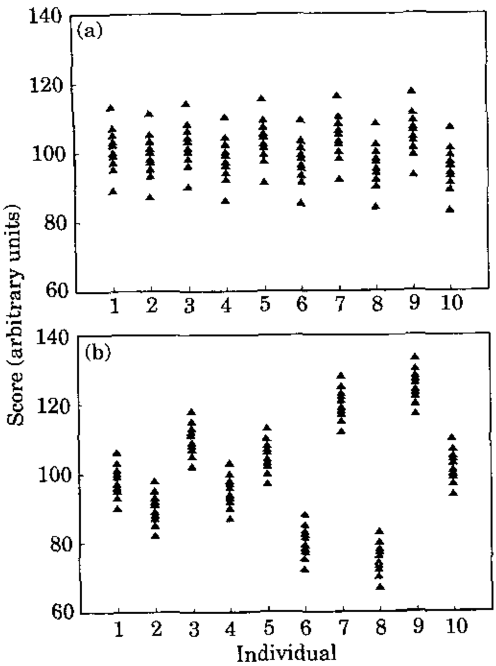

Figure 1. Distributions of scores from two simulated data sets. Populations 1 (a) and 2 (b) have intra- to between-individual variance ratios of 1·9:1 and 0·3:1, respectively.

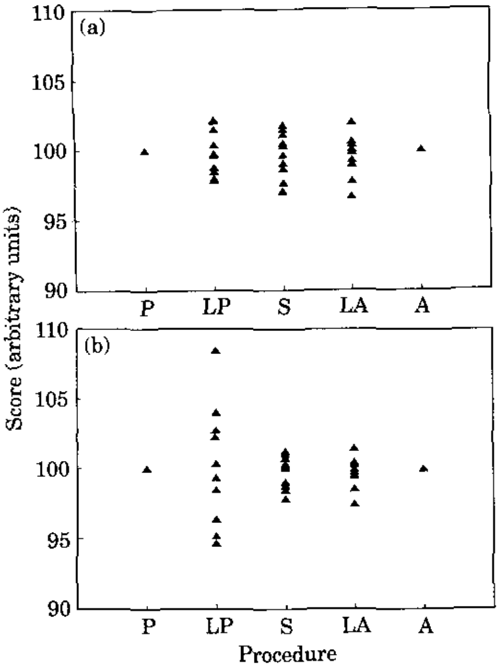

Figure 2. Estimates of the population mean for Population 1 (a) and Population 2 (b) made by five different procedures. P: Complete pooling; LP: limited pooling; S: single sampling; LA: limited aggregation; A: complete aggregation.

one score, but when an individual was sampled twice, we computed the mean of its scores. We imposed an upper limit of two scores per individual. In the 10 replications of this procedure, the number of individuals with two scores ranged from three to five. (5) Complete aggregation: we computed the mean of all scores for each individual and used these means to estimate the population mean and standard deviation. Note that in all procedures except complete pooling, we used 10 data points to estimate the population mean and standard deviation, although in both aggregation procedures we derived some or all of these 10 data points by averaging.

Results

Figure 2 illustrates the results of the sampling procedures for estimating the population means. In Population 1 (high intra-subject to between-subject variance ratio) there was no systematic relationship between sampling procedure and estimates of the mean. Limited pooling, single-sampling, and limited aggregation, each replicated 10 times, produced ranges of estimates that approximated the population mean. Furthermore, the range of the estimates was about the same in all three procedures, suggesting equal reliability.

In Population 2 (low intra-subject to between-subject variance ratio), the results were similar to those in Population 1, but with one important difference. Limited pooling resulted in extremely diverse estimates of the mean. When individuals differ substantially from one another and some individuals have more scores included in the sample, the potential exists for markedly biasing the outcome. Thus, pooling is unreliable when the population of scores is drawn from individuals whose means are quite different from one another and when those individuals are represented by unequal numbers of scores.

Estimates of the populations' standard deviations are shown in Fig. 3. In Population 1, all procedures except complete aggregation produced estimates that were close to the population standard deviation of 7·1. Complete aggregation produced a substantially lower estimate because it

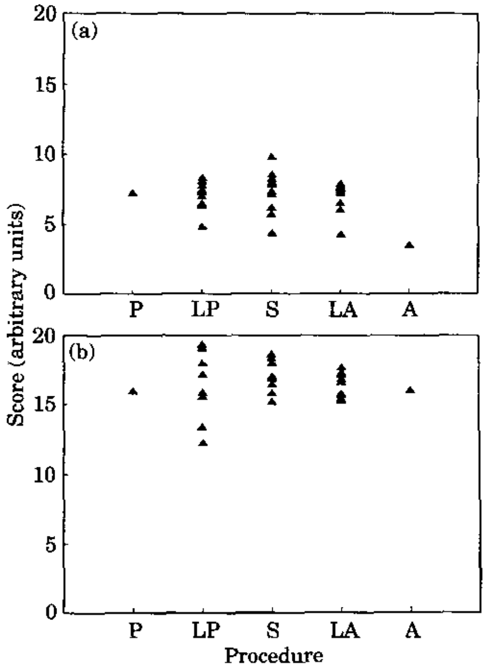

Figure 3. Estimates of the population standard deviation for Population 1 (a) and Population 2 (b) made by five different procedures. Procedure abbreviations are the same as in Fig. 2.

used the mean of all scores of all individuals, making it an estimate of between-subject variability. All other procedures used at least some raw scores, making them, to varying degrees, estimates of intra-subject variability as well. In Population 2, limited pooling led to widely varying estimates of population standard deviation. Again, with unequal subject representation in the samples (both in terms of unequal *N*s within a replicate and different subsets of individuals between replicates), the estimates of variability were relatively inconsistent from one replication to the next, even though their mean was approximately the same as those of the other procedures. (Figure 3a and b are drawn to the same scale to facilitate comparisons. Population 1 was less variable overall than Population 2, which accounts for the different placement of data points on the vertical axes.)

STUDY 2: MARMOT ALARM CALLS

The analysis of two simulated data sets suggests that pooling with unequal *N*s may produce

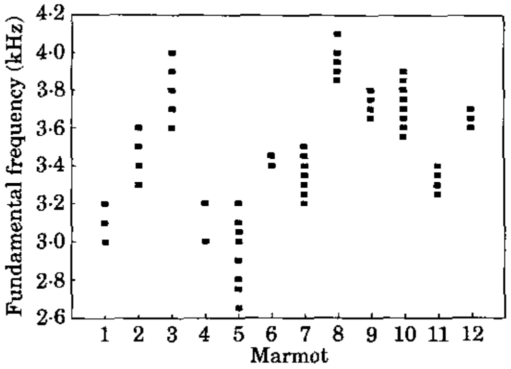

Figure 4. Distributions of fundamental frequency data on 220 marmot alarm calls from 12 individuals. Many data points overlap. Intra- to between-individual variance ratio was equal to 0·25:1.

inconsistent results when intra-subject variance is small relative to between-subject variance. We now turn to an actual data set that has that characteristic.

Methods

We recorded alarm calls of yellow-bellied marmots, *Marmota flaviventris*, on a Uher 4200 recorder using a Sennheiser ME-88 microphone. We worked at the Rocky Mountain Biological Laboratory near Crested Butte, Colorado. All calls were apparently in response to our approach or proximity.

Recordings were digitized at a sampling frequency of 20 000 Hz using a Personal Acoustics Laboratory (PAL) system (Davis 1986). The digitized data were used to measure the fundamental frequency of each call.

Over 350 calls were recorded, but our analysis is confined to those calls recorded from known individuals. There were 220 such calls from 12 animals (mean=18·3 calls/animal; range=5–36 calls). The data set is shown in Fig. 4.

We drew samples from the set of 220 calls in the following ways. First, in complete pooling, we used all 220 cases independently. In limited pooling, we randomly selected three calls per individual. In single sampling, we randomly chose one call per individual. In limited aggregation, we randomly selected three calls per animal but computed a mean of those scores before proceeding. Finally, in complete aggregation, we computed a mean for each animal's complete data set. Limited

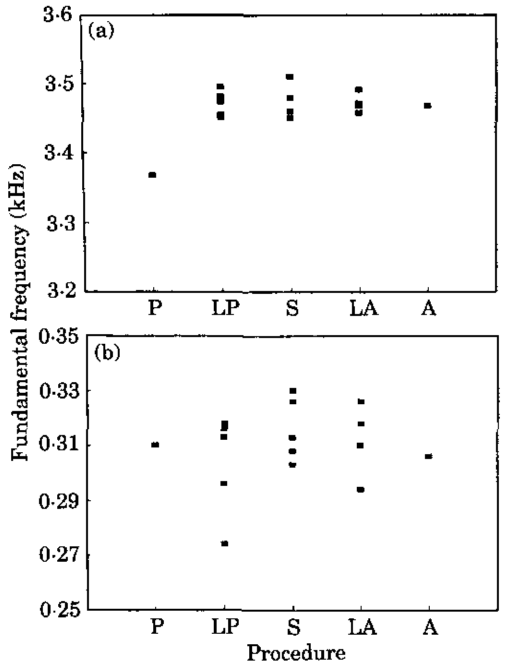

Figure 5. Estimates of mean (a) and standard deviation (b) of marmot alarm call fundamental frequency. Procedure abbreviations are the same as in Fig. 2.

pooling, single sampling, and limited aggregation were replicated five times each.

Results

The intra-subject to between-subject variance ratio in this data set was 0·25:1. Complete pooling yielded a conspicuously lower estimate of the mean than those obtained with the other procedures (Fig. 5). This occurred because the individual with the largest number of calls also had a low fundamental frequency. This is potentially a problem whenever there are unequal *N*s and a small intra-subject to between-subject variance ratio. Limited pooling in this case was not affected because all animals were represented by the same number of scores. In Population 2 of the simulated data, which had a similar intra-subject to between-subject variance ratio, the completely pooled data set had equal *N*s among subjects, but the limited pooling samples had unequal *N*s. Clearly, equal *N*s among subjects is important when pooling in data sets with this variance ratio.

Estimates of standard deviation were not strongly affected by data analysis procedure (Fig. 5). Because there was so little intra-subject variance, there was relatively little reduction in between-subject variance when data aggregation was performed.

STUDY 3: FORAGING TIME IN GROUND SQUIRRELS

Methods

We analysed data on time allocation in California ground squirrels, *Spermophilus beecheyi*, originally published by Leger et al. (1983). This study is typical of those in which pooling occurs: data were collected from a blind on a population of individually marked animals. Some individuals were measured only once, but others were measured repeatedly.

Data were collected on 31 individuals over a period of 3 months. (The original report mistakenly stated that there were 34 animals. This was due to miscoding of one animal's dye mark.) Each observational session consisted of 20 instantaneous samples (Altmann 1974; Leger 1977) spaced 30 s apart for 10 min during which the focal animal's behaviour category was recorded. Although several behavioural categories were used, we here confine our analysis to foraging behaviour, which was by far the most common behaviour category. Each session yielded a single value, the percentage of samples that were of foraging. Additional details about the procedure are presented in Leger et al. (1983).

There were a total of 116 sessions, or a mean of 3·74 sessions per squirrel. However, the number of sessions per animal ranged from one to 10 (Fig. 6). We analysed the data as follows. (*1*) Complete pooling: we analysed all the session scores ignoring the fact that many of them were repeats on some individuals. (2) Limited pooling: we randomly selected up to three scores per squirrel, but nine animals had only one score, seven animals had only two scores, and two other animals had only three scores each. (3) Single sampling: we randomly selected one score per individual (for the 22 animals that had more than one session). (4) Limited aggregation: we averaged up to three randomly selected scores per individual. (5) Complete aggregation: we averaged all the scores for each animal, regardless of how many scores there were. Limited pooling, single sampling, and limited aggregation procedures were replicated five times each.

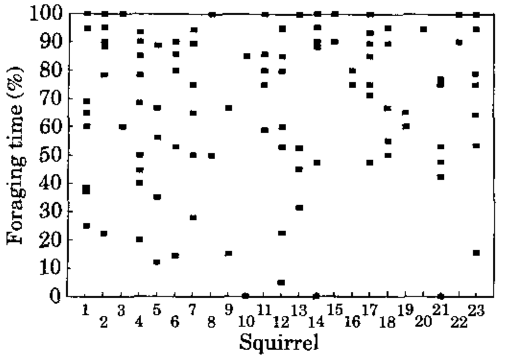

Figure 6. Distributions of foraging time data from 116 sessions from 31 ground squirrels. The nine animals with only one datum each are grouped together as squirrel 1. Intra- to between-individual variance ratio was equal to 1·03:1.

Results

The results are presented in Fig. 7. This data set had an intra-subject to between-subject variance ratio of 1·03:1. Single sampling produced the greatest range in estimated means, as one would expect given the large intra-individual variation. In fact, the range would have been even greater had there been fewer animals with only one score each, because their scores remained constant across the five replications. Limited pooling was more consistent than single sampling because, in addition to the nine animals with only one score each, nine others had two or three scores each, which were constant across replications. Limited aggregation yielded the most consistent estimates because extreme scores tended to regress towards the individual's mean when averaged with other scores from the same individual. Thus, when there is large intra-subject variance, single sampling may produce highly varied results. The same is true, although to a lesser degree, with pooling.

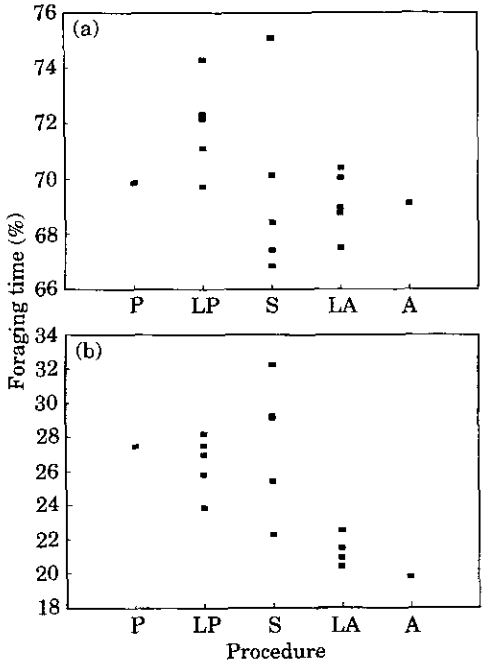

Figure 7. Estimates of mean (a) and standard deviation (b) of ground squirrel foraging time. Procedure abbreviations are the same as in Fig. 2.

Estimates of variability paralleled those of the mean. With single sampling, standard deviation scores varied substantially. Estimates derived by the use of limited pooling were less varied. Limited aggregation yielded standard deviations that were lower and less varied than those of pooling and single sampling. As squirrels were represented by data points derived by averaging two or three scores, their individual differences were reduced as one would expect given that individual means were about as different from each other as were scores from the same individual.

STUDY 4: HUMAN INFANT CRY ACOUSTICS

Machlis et al. (1985) claim that pooling has its most important effects when groups are compared. Because of the larger number of cases obtained through pooling, degrees of freedom increase. For instance, if 20 individuals from each of two groups are sampled five times each, the degrees of freedom for the error term would be 198 in the case of pooling, but would be only 38 if the data were aggregated by computing a mean for each individual. One should note, however, that although pooling does indeed increase the chance of rejecting the null hypothesis, because of the rapidly asymptotic nature of the F-distribution, the increase would be negligible (in the case of 38 and 198 degrees of freedom, the critical values differ by only 0·19 at alpha=0·05). Choosing a more conservative critical value of F would more than compensate for the greater degrees of freedom obtained by pooling. We focus now on the effects of pooling and aggregation on tests of differences between groups.

Methods

The data presented here are drawn from a study conducted in our laboratory. Briefly, we recorded the spontaneous cries of 20 infants (10 1-month-old and 10 6-month-old) during day-long recording sessions in the infants' homes. We recorded 250 crying episodes (140 from the 1-month-olds and 110 from the 6-month-olds). The mean number of crying episodes per infant was 12·5 (range=4–25).

We digitized up to the first 45 s of each recorded episode on a Personal Acoustics Laboratory system (Davis 1986). We measured or calculated 26 acoustic variables. We tested for age differences on all variables using analysis of variance based on completely pooled and completely aggregated data.

Results

The pooled and aggregated analyses agreed on 18 of the 26 (69·2%) acoustic variables on a simple classification into those which had significant versus non-significant age differences. Seven of these 18 variables yielded significant ($P<0·05$) differences between the age groups when pooled and when aggregated. The other 11 variables were not significant according to either procedure. Of the eight variables on which the analyses disagreed, seven had significant age differences when pooled, but not when aggregated. This is to be expected because of the greater degrees of freedom associated with the pooled data set. The last of the eight 'disagreement' variables was significant when aggregated but not when pooled. This outcome, which was nearly achieved in one other variable, was not expected given the large difference in degrees of freedom. However, it suggests a situation in which aggregation may produce a non-conservative test of the null hypothesis, and therefore will be discussed in more detail.

This variable, the standard deviation of the pause duration following 'fusses', was not significant when pooled ($F_{1,209}=2·83$, $P>0·05$), but was significant when aggregated ($F_{1,18}=4·76$, $P<0·05$). The omega-squared (Keppel & Sauffley 1980) values were 0·0086 and 0·1582, respectively, indicating that age accounted for about 18 times more variance when scores were aggregated than when they were pooled.

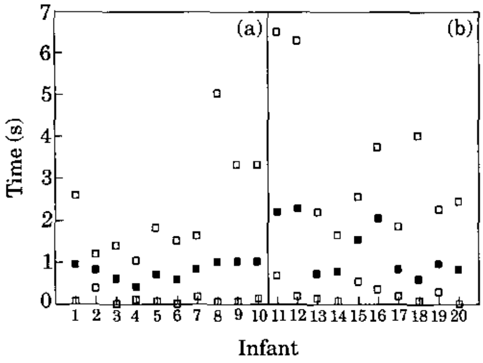

Figure 8. Distributions of inter-fuss interval variability in cries at (a) 1 month and (b) 6 months of age. □: Highest and lowest scores for each infant. ■: infants' means.

This result appears to be due to small differences between subject means (within age groups) combined with rather large intra-subject variation. This was particularly true in the data for infants at 1 month of age in which between-subject variance (based on completely aggregated scores) was only about one-third as great as the mean intra-subject variance (Fig. 8). Aggregation reduced within-group variance by eliminating intra-subject variance. Although the difference between the two age group means was quite small (0·79 and 1·28, respectively), the reduction in within-group variance made this difference significant.

The effect described above can be readily seen in a small set of hypothetical data (Fig. 9) consisting of two groups of three subjects, each of whom has four data points. When treated in their pooled

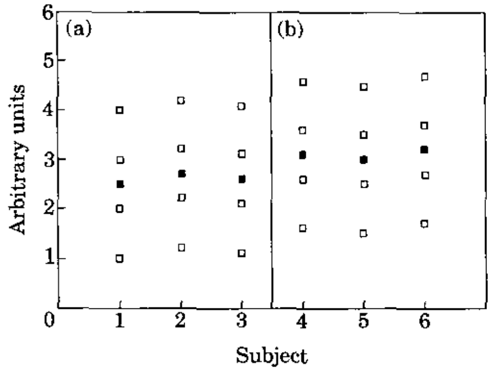

Figure 9. Hypothetical data on two groups (a and b) of three subjects. □: Data points (N=4 each); ■: subject means. The two groups differed significantly when data were aggregated but not when pooled.

form, the difference between the two groups was not significant ($F_{1,22}=1\cdot09$, $P>0\cdot05$) but when aggregated, the two groups were highly significantly different ($F_{1,4}=37\cdot5$, $P\ll0\cdot01$). The intra-subject to between-subject variance ratio in this case was 12·9:1, suggesting that aggregation may not be an appropriate procedure for testing group differences when intra-subject variance exceeds between-subject variance.

One could argue that the intra-subject variance should be eliminated by aggregating data if one wants to test for group differences. We would argue that if the independent (or grouping) variable is important, its effect should appear through intra-subject variance as well as within-group variance based on subjects' mean scores. According to this perspective, pooling data would produce a more conservative test of the age-difference hypothesis than would data aggregation, provided that intra-subject variance exceeds between-subject variance.

DISCUSSION

Our results indicate that pooling provides estimates of population means and variances that are at least as reliable as those provided by single sampling and aggregation, provided that the number of scores obtained from one's subjects is equal or that intra-subject variance is greater than between-subject variance. In one simulated data set in which intra-subject variance exceeded between-subject variance, pooled data provided reliable estimates of the mean when the number of scores per individual varied and even when different subsets of individuals were drawn for analysis. When intra-subject variance is less than between-subject variance, unequal *N*s become problematic. In a second simulated data set, which had this sort of variance profile, unequal *N*s produced unreliable results when pooled. In another data set with this profile (on marmot alarm call characteristics), pooling was unreliable when unequal *N*s were used but were reliable with equal *N*s. Thus, contrary to some claims made by Machlis et al. (1985), pooling does not necessarily invalidate the conclusions drawn by researchers who use it. Moreover, we have suggested a simple metric that researchers can apply to evaluate the appropriateness of pooling in their data sets.

Because testing hypotheses about group differences involves both between- and within-group variances, the effect of pooling on the outcome will depend, in part, on the ratio of intra-subject variance to within-group variance. If adding additional scores to the analysis increases within-group variance (on average), then doing so is conservative even if such scores come from previously measured individuals. Every data point increases the degrees of freedom for the error term, but the effect on the mean-square error depends on the magnitude of the score's deviation from the group mean. Pooling may be a conservative procedure, provided that successive scores from the same individual are as likely to deviate as much from each other as their mean does from the mean scores of other individuals. Despite this claim, relying too heavily on pooling from a small number of subjects may mean that the population of individuals has not been adequately sampled.

Avoidance of pooling does not guarantee valid results. Our analyses have shown that the two major alternatives to pooling have their own problems. Single sampling of individuals tends to produce less consistent estimates of means and variances (owing to sampling error), and aggregation may reduce estimates of between-subject variability and can increase the chance of making type I errors in some cases.

The decision to pool data or not is independent of decisions regarding procedures for sampling individuals from the population. Subject sampling is an important issue, and can be potentially troublesome in field conditions in which one has relatively little control over the comings and goings of potential subjects. Altmann (1974) does an excellent job of warning about potential biases that may occur. However, whenever more than one data point is collected per individual (for either pooling or aggregation), one must be concerned about the possibility of biases in intra-subject sampling. For instance, if the song rate of one bird is measured several times at a particular perch site, and another individual is measured several times at another site, one may have a biased sample of behaviour from both individuals, even though there may be an unbiased sample of individuals from the population. It may be inappropriate to pool or aggregate under such circumstances. Specific mention should be made of the conditions in which repeated samples are drawn from individuals, regardless of whether these samples are pooled or aggregated.

Extensive pooling, however, can mean that conclusions might be based on rather small samples of individuals, with the dangers inherent in such policies (Martin & Kraemer 1987). The danger, of course, is that the small sample is not representative of the population. However, a large sample can also be non-representative. We suggest that sampling procedures should be the main focus of the evaluation of research. The use of pooling should probably be of secondary concern. In other words, if one acquires a representative sample of individuals and a representative sample of their pattern of behaviour, pooling is a valid procedure provided that intra-subject variance *exceeds between-subject variance or that subject* *N*s are equal.

Our findings are relevant to methodological decisions made by animal behaviourists. These decisions occur at two times: during data collection and during data analysis. During data collection, researchers may be faced with the following problem: should a second (or subsequent) sample be taken from an individual that has already been measured, or should the researcher collect data from a previously unmeasured individual? Generically, the question concerns the issue of unequal *N*s, since precisely the same problem occurs when individual A has been measured, let's say, five times and individual B has been measured only twice. Our advice is to always measure the individual that has been measured least often in the past. If the choice is between re-measuring an individual versus waiting for an unmeasured individual, the decision hinges on how much time one might have to wait, how expensive the measurement process is, and so on. Provided data collection is not too expensive in terms of time or cost, obtaining additional data on previously measured individuals is a wise decision. More data, even if pooled, is more valuable than less data.

The second sort of methodological decision occurs during data analysis. The researcher can deal with multiple data points on individuals in three ways: by pooling them, by randomly selecting one datum per individual, or by aggregating. If pooling or aggregating, a decision can be made about how many data points to include from individuals. We suggest that such decisions be informed by the intra-subject to between-subject variance ratio. If this is high, even unequal *N*s seem to do little violence to estimates of the mean and variance. If it is low, pooling may be appropriate provided one uses the same number of data points per individual. With more than one data point per individual available for analysis, multiple estimates of the population mean and variance are possible, which permits one to scrutinize the reliability of the estimate.

Finally, the more one knows about the characteristics of the population being studied, the better informed one's data collection decisions will be. Therefore, it may be appropriate, useful, and even necessary to devote some data collection effort to obtaining multiple samples from at least some individuals so that intra-subject to between-subject variance can be calculated.

In conclusion, the interpretation of data has been, and always will be, a subjective process, guided by the distributions of hypothesis-testing statistics and by our understanding of the procedures used to generate the data. Pooling is an effective procedure provided our conclusions about the data do not extend beyond its limits. In that sense, pooling is no different than any other procedure.

ACKNOWLEDGMENTS

We wish to thank Calvin P. Garbin, Jeffrey A. French, Susan Lingle and C. Ray Chandler for their comments on an early draft of this paper, and Al Kamil and Don Jensen for their insights on these issues. The ground squirrel data were collected during field work supported by NSF Grant BNS 78-17770 to Donald Owings; the marmot data were made possible by a grant from the Stanton Memorial Fund at UNL to Inta Didrichsons; and the human infant study was supported by a UNL Research Council Award to Daniel Leger and Ross Thompson. Some of these data were presented at the Animal Behavior Society meetings in Kingston, Ontario in June 1992.

REFERENCES

Altmann, J. 1974. Observational study of behavior: sampling methods. *Behaviour*, **49**, 227–265.

Beal, K. G. & Khamis, H. J. 1990. Statistical analysis of a problem data set: correlated observations. *Condor*, **92**, 248–251.

Davis, R. O. 1986. The Personal Acoustics Lab (PAL): a micro-computer-based system for digital signal acquisition, analysis, and synthesis. *Comp. Meth. Prog. Biomed.*, **23**, 199–210.

Hoekstra, J. A. & Jansen, J. 1986. Statistical significance in comparative ethological experiments. *Appl. Anim. Behav. Sci.*, **16,** 303–308.

Keppel, G. & Sauffley, W. H., Jr. 1980. *Introduction to Design and Analysis: A Student's Handbook.* San Francisco: W. H. Freeman.

Kroodsma, D. E. 1989. Suggested experimental designs for song playbacks. *Anim. Behav.*, **37,** 600–609.

Kroodsma, D. E. 1990. Using appropriate experimental designs for intended hypotheses in 'song' playbacks, with examples for testing effects of song repertoire sizes. *Anim. Behav.*, **40,** 1138–1150.

Leger, D. W. 1977. An empirical evaluation of instantaneous and one-zero sampling of chimpanzee behavior. *Primates*, **18,** 387–393.

Leger, D. W., Owings, D. H. & Coss, R. G. 1983. Behavioral ecology of time allocation in California ground squirrels (*Spermophilus beecheyi*): microhabitat effects. *J. comp. Psychol.*, **97,** 283–291.

Machlis, L., Dodd, P. W. D. & Fentress, J. C. 1985. The pooling fallacy: problems arising when individuals contribute more than one observation to the data set. *Z. Tierpsychol.*, **68,** 201–214.

Martin, P. & Kraemer, H. C. 1987. Individual differences in behaviour and their statistical consequences. *Anim. Behav.*, **35,** 1366–1375.