Studies in Avian Biology No. 6:388–389, 1981.

SUMMARIZING REMARKS: OBSERVER VARIABILITY

LYMAN L. MCDONALD[1]

In general there are four sources of "error" or variation in scientific studies (Cochran 1977):

(1) *Sampling error* due to inherent variability between experimental units. Thus, if a study area is divided into quadrats and each member of a sample is censused perfectly, summary statistics will necessarily vary from sample to sample. Another sample will yield another estimate due to sampling error.
(2) *Measurement error* due to the lack of uniformity in the physical conduct of the study. The measurement procedure may be biased, imprecise or both biased and imprecise.
(3) *Missing data* due to the failure to measure some units in the sample.
(4) *Gross errors* introduced in coding, tabulating, typing and editing the results.

Usually the effects of these errors are completely confounded and the total variance cannot be separated into its components. An understanding of sampling error and its role in making inductive inferences is the basis of modern statistical inference procedures. Control of this source of error is at least partially the responsibility of the statistician. Control of the other three sources of error is primarily the responsibility of the researcher! All of the papers in this session on observer variability have as a first objective the control of measurement error, and I applaud their attempts to get a handle on this problem. Measurement errors may be modeled by statisticians but their control and reduction must come from careful experimental design. Consultation between the researcher and statistician before the study begins should be of value in controlling all potential sources of error.

In many fields of study the presence of measurement error is barely recognized and its influence is played down. For example, Box et al. (1978) state that "Usually only a small part of it (the total variance) is directly attributable to error in measurement." Many statisticians follow the rule of thumb that the measurement error should be "small" relative to the sampling error, especially in utilizing statistical procedures such as regression and correlation analysis. Considering the content of the papers in this session, measurement errors cannot be ignored in studies designed to measure terrestrial bird numbers. Furthermore, standard analysis procedures may not be applicable until this source of error is under control.

Robbins and Stallcup (1981) consider a particular type of measurement error, namely inaccurate lists of species present at a study site. For instance, they mention a study in Maryland in which there was not one stop out of 50 at which two observers had recorded the same list of species present. They also briefly address errors in the fourth class, that is, errors occurring between the time a bird is observed and the time the report appears in print. Cyr (1981) reports on an experiment to test the ability of observers to identify species from utterances recorded on a tape. Scott et al. (1981b) report on experiments to study the ability of observers to estimate distances and the effect of bias in this process on their estimates. Emlen and DeJong (1981) propose to attack this problem by determination of detection threshold distances for each species under standard conditions. Supposedly these detection threshold distances could then be used as the half-width of transect censuses and the radius of point-centered census plots. Unfortunately, their proposal is still subject to measurement errors. Two observers running the same census plots or lines at the same time will have different counts of birds heard even though they might be willing to use the same detection threshold distance. To quote Kepler and Scott (1981), "Thus, an observer with good ears is actually sampling a larger area by hearing more distant birds."

Errors or variance due to missing data are not directly addressed in these papers. However, Cyr's (1981) experiment suffers somewhat from this source. Thirty-three observers apparently started the experiment but in the end only the results of eighteen were analyzed. There may be good reasons to drop the data from those fifteen observers but it is obvious that their retention would produce different summary statistics. That is, missing data is a source of variation in scientific studies. For example, if Cyr had been able to retain two "inexperienced observers" instead of only one, or if the single inexperienced observer had been dropped, the results of his regression analyses in Figure 2 would likely have changed drastically.

The following are mentioned as general procedures which may help to increase the precision of bird studies by decreasing sampling error:

[1] Departments of Statistics and Zoology, University of Wyoming, Laramie, Wyoming 82071.

1. Use careful *stratification* of the study site. Ideally, we would like to stratify on bird density itself, but in practice one must stratify on variables which hopefully are highly correlated with bird density. For example, in the Red Desert of Wyoming the researcher might stratify on the proportion of area "covered" by shrubby vegetation. Regions with "high" cover form one stratum, regions with "medium" cover form another stratum, etc. and each stratum is sampled independently.
2. Measure covariates on the sample plots (transects) which may help explain variation in bird density. Again, in the Red Desert of Wyoming the researcher might divide the study area into rectangular quadrats and estimate bird density (the variate) and cover by shrubby vegetation (a covariate) on each plot. Bird density estimates may show a significant reduction in variation when "adjusted" for the cover values in a regression analysis.
3. Use systematic or cluster sampling plans which may speed up the data collection and hence enable one to increase sample sizes.

Continuing the list for the control of measurement errors I would suggest:

4. Use of double sampling with ratio or regression estimators. For example, the accuracy of the variable circular plot survey (Ramsey and Scott 1979) might be enhanced by double sampling where distances to birds are measured on a subset of the sample and distances are estimated for all birds. Perhaps the measured distances could be used to "calibrate" the estimated distances.
5. Refinement of the experimental design (i.e., the physical conduct of the study). All of the papers in this session fall into this category. For instance, Robbins and Stallcup (1981) recommend "methods based on repeated visits over several days by different observers . . . ," careful training of observers and careful examination of field records. Kepler and Scott (1981) also stress the value of training observers. Faanes and Bystrak (1981) stress the importance of choosing well-trained observers whose differences will contribute little beyond sampling error.

A few specific remarks should be made concerning some of the papers. Scott et al. (1981b) are very careful to point out that the 20% error of estimation in their computer simulation was due to errors in measurement of distance and that no other sources of error were simulated. Factors such as observer bias will in some cases tend to counteract the measurement error and in other cases tend to magnify the error. It is important to remember that the 20% figure is for only one component of the many possible sources of error. Undoubtedly other sources will be incorporated into future simulations. Also, sample sizes in Scott and Ramsey's simulation were fairly large (i.e., 200–250 birds). A reduction in the sample size will likely result in increased error. They report approximately 10% accuracy in distance estimates from their field work, but this figure is for the *mean* accuracy while individual estimates varied from −75% to +400%. Again, to obtain this accuracy in estimation of the mean distance, sample sizes will have to be approximately equal to those employed in their field study.

Faanes and Bystrak (1981) have stated that "In most cases, well-trained observers are comparable in ability and their differences contribute little beyond sampling error." I think that they would agree that their sample of well-trained observers is fairly small, namely a sample of two consisting of the two authors. The same criticism of small sample sizes is valid throughout most of their paper. Their inductive inferences may remain valid, but one would like to see a broader sample from the population of observers. It is also dangerous to compare observers when they conduct the survey in different years. The year effect and observer effect are completely confounded, and the strength of the inference is decreased.

In conclusion, many of the problems facing researchers in the estimation of bird density deal with observer variability. Similar problems exist in finite sampling theory under the heading of "interviewer bias," see for example Cochran (1977). There is an extensive literature on the control, reduction, and evaluation of interviewer bias. Perhaps review of that literature will provide new ideas for research on observer variability in the estimation of the numbers of birds.