wise incorrect labels. The subjects often remarked on these incongruities and made it apparent that a far miss only rarely carried with it a conviction that it was the veridical label.

11. Lawless and Engen (*5*) have termed this failure to retrieve the label for an odor the tip-of-the-nose phenomenon.
12. Correlation between number of veridical labels emitted during inspection and subsequent overall performance equaled .73.
13. There remains the question of why undiscarded nonveridical labels showed any serviceability whatsoever. The relatively high variability of the performance obtained with nonveridical labels (Fig. 2) reflects some of the reasons. Some nonveridical labels, even far misses, seemed to possess personal meaning (for example, Dad's bathroom) that endowed them with serviceability despite their surface imprecision. Other nonveridical labels apparently held neither personal meaning nor high surface precision and showed virtually no serviceability. In addition, even far misses often contained considerable generic information, as seen in the use of terms like spice for cinnamon, industrial chemical for turpentine, and so forth. Such generic terms led to erratic though hardly negligible identification. Finally, if a particular verbal-verbal link served well, the subject could merely choose to retain the nonveridical label [L. S. Prytulak, *Cognit. Psychol.* **2**, 1 (1971)]. Hence, nonveridical labels seemed to comprise a potpourri of personally meaningful, useless, and partially informative labels, and the high accompanying variability seems compatible with this diversity.
14. Abrupt jumps also accompanied switches from near misses to veridical labels (a factor of 2.2) and from far misses to near misses (a factor of 2.0).
15. Desor and Beauchamp (*4*) trained three subjects to identify almost 60 out of 64 odorants through use of a complex regimen of massed and distributed practice over many days using corrective feedback with veridical labels. The outcome of their experiment dispelled any doubt that laypersons could actually perform better than previously suspected, but left undetermined the reasons a training regimen succeeded.
16. Familiarity, though deliberately restricted here through the choice of common stimuli, played a role in all three of the present identification experiments. Average familiarity ratings correlated with subsequent identification in the following way: $r = .86$, .73, and .59 for experiments 1, 2, and 3, respectively ($P < .01$ throughout). Uncommon and hence unfamiliar stimuli would therefore seem to stand little chance of identification unless the subjects received long and arduous training. In fact, R. G. Davis [*J. Exp. Psychol. Hum. Learn. Mem.* **104**, 134 (1975)] found that subjects failed even to approach perfection after 20 trials when required to identify only four relatively unfamiliar odorants with numerals.
17. E. R. Bickerstaff, *Neurological Examination in Clinical Practice* (Blackwell, Oxford, 1968), p. 36.
18. W. S. Cain and J. Krause, *Neurol. Res.*, in press.
19. Jones (*2*), who permitted two perfumers to choose their own stimuli (perfume ingredients) and then tested identification with small sets of the chosen stimuli, estimated that such professionals could probably identify 100 to 200 odorants. Because the stimulus sets apparently excluded substances commonly encountered in the everyday lives of most laypersons, however, Jones's estimate may represent the increment that one type of professional experience can add to the relatively large number of common substances that laypersons can identify under the right circumstances.
20. The data depicted by the unfilled circles represent weighted averages. Hence, the breakdown of performance by the quality of label gave subjects who emitted more than average veridical labels a heavier weight in the calculation of the function for veridical labels and a lighter weight in the calculation of one or both of the other functions. As it turned out, this factor had virtually no net influence on the functions for veridical labels and near-miss labels. If all subjects were given equal weight in the final tally, irrespective of how many labels of each type they had emitted, these two functions would differ from those shown in the figure by less than 1 percent. On the other hand, the function for near-miss labels would rise by about 9 percent. The rise would reflect the increase in the relative contribution of those high-scoring subjects, noted in the text, who emitted fewer than average far misses but who could use virtually any label effectively. Nevertheless, the finding that the manner of computation would not change the functions for the veridical labels and near misses shows that the high performance obtained in these cases did not represent a statistical segregation of generally high-scoring subjects from low-scoring ones.
21. Supported by NIH grant ES-00592. This investigation began as a senior-year research project by R. Sax of Yale College. I thank him for his efforts in the early stages and H. G. Anderson III and R. J. Huey for technical assistance.

21 August 1978

Regularity, Randomness, and Aggregation in Flowering Phenologies

Stiles (1) presented the results of a 4-year study on the flowering times of 11 hummingbird-pollinated plants in a Costa Rican rain forest. Stiles was primarily concerned with testing the hypothesis that "a system of compensating phenological responses of different species to unusual rainfall conditions may play a major role in maintaining an orderly, staggered sequence of flowering peaks among the hummingbird-pollinated plants." The basis of this hypothesis was the belief that natural selection should produce a regular sequence of flowering times, in order to minimize competition between plant species for pollinating hummingbirds or to minimize interspecific hybridization. Stiles concluded that "The phenological data . . . show that a regular sequence of flowering peaks was nearly always maintained . . . only during late November to early December was no hermit food plant ever at peak bloom. . . ."

The crux of Stiles' argument lies in demonstrating that the flowering times shown in his figure 1 are indeed regularly spaced. Stiles' conclusion that the pattern is regular within any one year is apparently based on a subjective examination of his data. One of us (B.J.R.) was faced with a similar situation in a study of flowering times in shrub communities. The flowering times observed in this study appeared regularly spaced, but unfortunately so did phenologies produced by assigning to each species a flowering time at random within the growing season. A subjective examination of the data was not sufficient to determine whether or not flowering times were, indeed, regularly, rather than randomly, spaced or even aggregated. Therefore, several statistical tests were developed to test the regularity hypothesis (*2*). One of these tests is applied below to Stiles' data. It shows that his sequences of flowering times in each of the 4 years are not regular as he concludes, but instead they tend to be aggregated in the drier parts of the year.

The null hypothesis is that the peak flowering date of each of the k species is independently and randomly assigned a position along an axis representing the growing season from a rectangular (uniform) probability distribution. The length of the growing season is then normalized to one for computational simplicity (each peak flowering date is divided by the length of the growing season). These randomly assigned flowering peaks $x_1, x_2, \ldots x_k$ are then ordered from earliest to latest, designated as the order statistics of the sample $y_1, y_2, \ldots y_k$ where y_1 is the earliest flowering species and y_k the last flowering species. The interval $y_{i+1} - y_i$ is then the distance in time between the peak flowering dates of any two adjacent flowering species. The null hypothesis is equivalent to the procedure of assigning to each of the k species a peak flowering date at random from a table of random numbers, ordering the random numbers from first to last, and then normalizing everything to one. Given the null hypothesis, the statistical properties of $y_{i+1} - y_i$ can be derived (*2*). In particular, the mean of $y_{i+1} - y_i$ is $1/(k + 1)$, and the variance is $k/[(k + 1)^2(k + 2)]$.

Consider the sample statistic P.

$$P = \frac{\sum_{i=0}^{k} \{y_{i+1} - y_i - [1/(k+1)]\}^2}{k+1}$$

which is the sample variance of the distances between peak flowering dates between adjacent species, including the distance between the beginning of the growing season and the peak flowering date of the first species to flower and between the last peak flowering date and the end of the growing season. The expected value of P under the null hypothesis of randomly assigned peak flowering dates is (*2*)

$$E(P) = \frac{k}{(k+1)^2(k+2)} \qquad (1)$$

If peak flowering times tend to be regularly distributed through the growing season, the sample variance P should be less than that expected from Eq. 1 (with 0.0 as a lower limit for perfect regularity); at the same time, if peak flowering dates are aggregated, the sample variance will exceed its expected value under the null hypothesis. The ratio $P/E(P)$ is, therefore, a measure of regularity or aggregation, and the chi-square statistic $\chi^2 = kP/E(P)$ with k degrees of freedom is an approximate test of the null hypothesis, that is, $P = E(P)$.

Some conventions are needed to apply this test to Stiles' data. First we chose the median or peak flowering date for each species as the midpoint of the line that Stiles uses to represent "peak bloom." Because the growing season is continuous in Costa Rica, we defined the growing season to be from the median flowering time of the first species as the beginning of the interval and the median flowering time of the last species of the year as the end of the interval. This convention eliminates the period from about the end of October to the middle of January, during which one species at most was in bloom because Stiles implies that this period may be an exception to his regular sequence of flowering. If this period is included in the analysis, the results given below are considerably more significant than they are if this more liberal convention is used.

Given these two conventions, there are nine intervals between median flowering times to be considered ($k = 8$). There are ten species in each of the 4 years from 1971 to 1974. The expected value of P with $k = 8$ is 0.0098. The ratios $P/E(P)$ for each of the 4 years are 2.0720 (1971), 1.8459 (1972), 2.0546 (1973), and 2.0264 (1974). That all four ratios are greater than 1.0 implies aggregation of flowering times. If the chi-square statistic with eight degrees of freedom is used, all 4 years are significantly more aggregated than would be expected under the random null hypothesis at the 10 percent level and two of the years approach significance at the 5 percent level. If the period from October to January is included in the analysis, all 4 years are significantly aggregated at the 5 percent level.

The analysis, therefore, does not support Stiles's conclusion that the flowering peaks of hermit hummingbird-pollinated plants are regularly spaced throughout the growing season. Instead, there is strong evidence that the flowering times are aggregated or at most randomly spaced. The aggregation exists apparently because the majority of the flowering peaks occur during the two drier periods of the year.

We wish to emphasize, however, that the analysis does not imply that competition does not exist or has not been a selective force on flowering times, only that Stiles's data are not regularly spaced and do not support the competition hypothesis.

ROBERT W. POOLE
Division of Biology and Medicine, Brown University, Providence, Rhode Island 02912

BEVERLY J. RATHCKE
Department of Ecology, Evolution, and Behavior, University of Michigan, Ann Arbor 48109

References

1. F. G. Stiles, *Science* **198**, 1177 (1977).
2. R. W. Poole and B. J. Rathcke, unpublished data.

21 February 1978; revised 6 October 1978

The comments by Poole and Rathcke (*1*) demonstrate that statistics is a two-edged sword when dealing with complex biological phenomena. Their statistical test is unfortunately based on a biologically unrealistic assumption, that of a "continuous growing season in Costa Rica." The very concept of a "growing season," derived as it is from temperate-zone agricultural practices, may be very inappropriate at the community level, and not only in the tropics: the actual growth of many temperate-zone woody plants occurs within the span of a very few weeks, rather than months (*2*). In fact, growth—and flowering—of most tropical plants is highly seasonal, and certain times of year in tropical habitats may be unsuitable or stressful for flowering (*3*). Competition for pollinators should tend to spread blooming peaks toward a uniform distribution over time, but the extent to which such a distribution can be achieved will be limited by the physiological constraints imposed upon the plants by the seasonal regime of the habitat. In a seasonally varying tropical habitat, a mathematically uniform distribution of flowering peaks may be impossible for real plants to attain.

My data (*4*) show why a uniform spacing of flowering peaks was not indicated by Poole and Rathcke's test. There are two clusters of such peaks, corresponding to the dry and early wet seasons, with only one species regularly attaining peak bloom in the interim. This pattern is, moreover, characteristic of the hummingbird-flower community as a whole and suggests that the dry-wet transition period is in fact unfavorable for flowering (*3*). This is in accord with the observation that wet-season and dry-season bloomers may respond oppositely to a given rainfall event, and tend to differ in habitat or growth habit (or both) as well (*3*). The kinds of physiological and ecological adaptations required for flowering in these two seasons might be qualitatively different, and the dry-wet transition might be favorable for neither group. In terms of the Poole-Rathcke model, the probability of a flowering peak falling at random in this period will be less than the corresponding probabilities for the dry and early wet seasons, and a biologically meaningful statistical test should take this into account. We require a base line of equal feasibility of flowering, in terms of the capacities of the plants, against which to test whether flowering peaks are more or less uniformly distributed through time. The problem lies in deriving such a base line without circular reasoning, especially in view of our present ignorance of the physiology of tropical plants.

A possible approach might be to use the distribution of flowering peaks of a large random sample of plants of similar growth habits in the community, as a "bioassay" of "flavorability" of different times of year for flowering—analogous to the procedure of Colwell and Futuyma (*5*) for evaluating niche differences. However, flowering in ornithophilous plants may be more costly with respect to producing, displaying, and protecting large quantities of nectar, than flowering in entomophilous species (*6*), and physiological constraints on flowering also might be more severe. In any case, I do not have this sort of information available from my La Selva study. At this stage our understanding of tropical flowering phenology can best be advanced by studies on the physiological ecology of the plants, rather than statistical tests, however sophisticated, that are based on clichés like the "continuous tropical growing season."

F. GARY STILES
Escuela de Biología, Universidad de Costa Rica, Ciudad Universitaria, Costa Rica

References and Notes

1. R. W. Poole and B. J. Rathcke, *Science* **203**, 470 (1979).
2. R. Borchert, *Physiol. Plant.* **35**, 152 (1975).
3. See F. G. Stiles [*Biotropica*, **10**, 194 (1978)] for data and literature review.
4. F. G. Stiles, *Science* **198**, 1177 (1977).
5. R. K. Colwell and D. J. Futuyma, *Ecology* **52**, 567 (1971).
6. F. G. Stiles, *Am. Zool.* **18**, 603 (1978).

21 November 1978