main topic of the paper. We regret any confusion that was created by our use of the term "K-feldspar" instead of x-ray antiperthitic alkali feldspar.

T. E. BUNCH
Space Sciences Division, Ames Research Center, National Aeronautics and Space Administration, Moffett Field, California 94035
EDWARD OLSEN
Department of Geology, Field Museum of Natural History, Chicago, Illinois 60605

References

1. G. J. Wasserburg, H. G. Sanz, A. E. Bence, *Science* **161**, 684 (1968).
2. T. E. Bunch and E. Olsen, *ibid.* **160**, 1223 (1968).

25 October 1968

Chromosomal Effect and LSD: Samples of Four

The analysis by Sparkes, Melnyk, and Bozzetti (*1*) on the effect of LSD in vivo on human chromosomes creates a misimpression, primarily because they neglect the effect of their very small sample sizes. Closely associated with the problem of sample size is their neglect of the distinction between statistical significance and substantive significance. The distinction, which has been made for years (*2*), still is frequently misunderstood.

Sparkes, Melnyk, and Bozzetti worked with three groups of four people each: controls, users of LSD, and people medically treated with LSD. About 225 lymphocytes from each of the 12 persons were examined, and a variety of kinds of chromosomal damage was observed. Four scoring schemes were used; for brevity we repeat here in Table 1 the results for only one. Then the Wilcoxon-Mann-Whitney test was repeatedly applied, and no statistically significant results (at the usual levels) were obtained.

Our major comment is that, in comparing two samples of size four, the substantive, real difference must be very large to have reasonable power, that is, to have a reasonably large probability of detecting the real difference. Therefore, a finding of no statistically significant difference does not by any means preclude the existence of a material real difference.

Before considering the question of power, we first mention another way of looking at the consequences of small sample sizes: confidence intervals. Let us assume that the sampled populations differ essentially only by translation; for example, let us assume that for some unknown number Δ the underlying distribution of cell percentages for users is the same as that for controls after adding Δ to each control percentage. Then the Wilcoxon-Mann-Whitney test is readily applied (*3*) to obtain confidence intervals for Δ. The results, at the 94.3 percent level of confidence (*4*), in percentage units, are (i) users minus controls, −5.5 to 6.7; (ii) medically treated minus controls, −3.3 to 2.3; (iii) users minus medically treated, −4.8 to 7.8. The first of the above, for example, says that the observed difference between users and controls is not surprising (5.7 percent significance level) if one were testing null hypotheses that the real difference lies between −5.5 and 6.7. It seems to us that real differences of 6 or 7 in percentage units might be quite important; such real differences are consistent with the observed data.

Similar conclusions are reached from the viewpoint of power. For example, if breakage-gap scores had negative exponential distributions, and a significance level of .057 were used, the null hypothesis would be rejected only about 60 percent of the time, even if users had an average breakage rate *six* times that of controls. For a significance level of .029, a corresponding percentage is only achieved with an average rate for users nine times that for controls (*5*). If the parent populations are normally distributed with common variance σ^2, it is notable that, for a significance level of .029, the probability of rejecting the null hypothesis for a difference in means of 2.5 σ is .682 (*6*). It may be seen from Table 1 that σ is quite substantial.

There are other difficulties in reaching conclusions from this set of data, and we mention three of them. First, there is no reason to think that the three samples are either random or from the same population of humans. Some differences are immediate; for example, the medically treated subjects range in age from 28 to 45, while the users age range is 19 to 24. The control ages go from 21 to 50. This problem of basic noncomparability may be inherent in studies of this kind, and we do not take the view that valid conclusions in such circumstances are impossible. Nonetheless, an extra measure of caution is necessary. Second, we do not know whether the cells were observed blindly, that is, with the observers in ignorance of the source of the samples. Since determination of cell aberration doubtless has some subjective elements, a lack of blindness might introduce bias. Third, two laboratories analyzed separate samples of blood from each person. The differences between the results from the two laboratories (which used different techniques) would be illuminating, since they would give an idea of variability stemming from both the blood sampling and from the laboratory techniques. Unfortunately, the only information given is that there was no significant difference between results from the two laboratories.

Table 1. Percentages of cells with breaks or gaps (*1*). Samples are ordered within themselves.

Sample	Broken cells (%)			
Controls	3.3,	4.8,	6.4,	7.1*
Users	0.9,	2.6,	3.4,	11.5
Medically treated	3.1,	3.7,	5.7,	7.1*

* The unrounded 7.1 for controls is slightly less than the one for medically treated.

WILLIAM H. KRUSKAL
SHELBY HABERMAN
Department of Statistics, University of Chicago, Chicago, Illinois 60637

References and Notes

1. R. S. Sparkes, J. Melnyk, L. P. Bozzetti, *Science* **160**, 1343 (1968).
2. E. G. Boring, *Psychol. Bull.* **15**, 335 (1919); W. H. Kruskal, in *International Encyclopedia of the Social Sciences*, D. L. Sills, Ed. (Macmillan and Free Press, New York, 1968), vol. 14, pp. 238–250.
3. L. Moses, in *Statistical Inference*, H. M. Walker and J. Lev, Eds. (Holt, New York, 1953), chap. 18.
4. Because the test statistic has a discrete distribution, it is difficult to use a conventional level like 95 percent. We chose the readily available level closest to 95 percent.
5. R. A. Shoruch, *Technometrics* **9**, 666 (1967).
6. W. J. Dixon, *Ann. Math. Statist.* **24**, 611 (1954).

12 September 1968

Sparkes *et al.* (*1*) tested the null hypothesis that there is no difference in chromosomal aberrations between users of LSD and nonusers. Their data indicated no significant difference in aberrations among test groups, hence the hypothesis was accepted. The probability level chosen in their test of significance specified the risk they were willing to take of rejecting the hypothesis if it were true (type I error). But there is also a risk of accepting a false hypothesis (type II error). The chance of making a type II error can be determined only for a specified difference between means. For any specified difference, however, it is possible to determine how many replications would be

necessary to keep the probabilities of making both kinds of errors at a certain level if one has an estimate of variability of the experimental material.

I have made this calculation using percentages of total aberrations (breaks plus gaps) (*2*). The percentages were transformed to the square roots of their arc sin values for obtaining an estimate of the variance. A normal distribution, necessary for the proper application of this method, is assumed arbitrarily. Six replications per group would be required to detect a mean difference of 5 percent per subject with a 5 percent risk of rejecting a true hypothesis and a 25 percent chance of accepting a false hypothesis. Only four replications were used in the experiment.

What this means is that the authors may be taking a greater than 1 in 4 chance of accepting the hypothesis that there is no difference between users and nonusers, even if a true difference of 5 percent total aberrations actually exists. I do not know whether the specified difference of 5 percent is appropriate; this is a medical question. It is, however, a rather large difference, with respect to the overall mean percentage, of 6.8 percent aberrations.

The importance of decisions based on the results of this experiment seems to warrant an attempt to reduce the risk of making a type II error by increasing replications. This is especially true in light of the fact that Sparkes *et al.* recognize that their results are at variance with other published work.

F. W. WHITMORE
Ohio Agricultural Research and Development Center, Wooster, Ohio

References

1. R. S. Sparkes, J. Melnyk, L. P. Bozetti, *Science* **160**, 1343 (1968).
2. G. W. Snedecor, *Statistical Methods* (Iowa State College Press, Ames, ed. 5, 1956), p. 275.

19 August 1968

We agree that, given the mean differences between control and LSD exposed subjects in our study, there is danger of a type II error when significant differences cannot be demonstrated and that since type I and type II errors are inversely related they can be minimized only by simultaneously increasing sample sizes. Situations in which the power of a test is reduced and the importance of both types of error are equal permit lowering of the accepted confidence limit (*1*). However, in the case of our combined results and counting all aberrations (breaks plus gaps), the null hypothesis could not be rejected even at the .20 level for control versus "users" and at the .10 level for the control versus LSD-treated subjects.

Further, it might be questioned whether the two types of error should be weighed equally. It might be argued that the acceptance of a fallacious hypothesis may be more detrimental to scientific progress than rejection of a true one.

We are unaware of any theoretical reason to anticipate whether LSD should have a damaging effect on chromosomes. Therefore, the answer to this important question has to be based on observations and there is no reason a priori for weighing levels of significance on the basis of what a reasonable result should be. We thus used the standard 5 percent confidence limit.

Because studies of chromosome damage often demonstrate a skewed distribution, with a few individuals showing a large number of aberrations relative to the rest of the sample, the use of a "distribution-free" test of significance, as applied in the evaluation of our data, seemed appropriate. Whether assumptions of normal distribution can be made, thus allowing for more powerful inferences, depends on one's judgment regarding the "robustness" of such procedures.

As Kruskal and Haberman note, complete random selection of subjects is difficult, and, the populations from which our three groups were drawn are different. Our greatest concern was with the exposure or lack of exposure to drugs. Second, the cells were evaluated blindly for chromosome damage, a point inadvertently omitted from our initial report. Third, portions of the same blood sample were analyzed in the two laboratories, and not "separate samples" as suggested by Kruskal and Haberman. The results from each laboratory were evaluated separately, results from each group of subjects were compared between laboratories, and then the results were combined. Comparisons between laboratories as noted in our Table 3 (*2*) indicate that the null hypothesis is sustained for the following P values: controls versus controls (breaks plus gaps), $P = .057$; controls versus controls (breaks), $P = .171$; users versus users (breaks plus gaps), $P = .557$; users versus users (breaks), $P = .443$; treated versus treated (breaks plus gaps), $P = .243$; and treated versus treated (breaks), $P = .443$. Results between groups in each laboratory were in the same direction for both laboratories.

With regard to the "substantive" significance of our findings in seven of the eight comparisons (*2*, Table 3) of controls with subjects exposed to LSD, the controls show a higher percentage of aberrations; the one exception is that in which controls had fewer breaks than the "users." Therefore, despite the above-mentioned limitations of the statistical evaluation of our data, we are still inclined to conclude that our studies do not show either "statistical" or "substantive" evidence of chromosomal damage by LSD.

ROBERT S. SPARKES
Department of Medicine, University of California School of Medicine, Los Angeles 90024
DAVID THOMAS
Department of Anthropology, University of California, Riverside 92502
JOHN MELNYK
Children's Hospital of Los Angeles, University of Southern California School of Medicine, Los Angeles 90033
LOUIS BOZZETTI
San Diego, California

References

1. B. V. Winer, *Statistical Principles in Experimental Design* (McGraw-Hill, New York, 1962).
2. R. S. Sparkes, J. Melnyk, L. Bozzetti, *Science* **160**, 1343 (1968).

9 October 1968

Factors Determining Spatial and Size-Frequency Distributions of Gemma gemma

Jackson (*1*) has used some data on the spatial and size-frequency distributions of *Gemma gemma* Totten in a bay near Guilford, Connecticut, to support his conclusion that "generalizations on the paleoecological significance of one sort of size-frequency distribution or another seem inappropriate without some idea of the life histories involved." Although we would not disagree with this conclusion, we feel, on the basis of our own work of the last 2 years at Barnstable Harbor, Massachusetts (*2*), that Jackson's data on *Gemma* and some of the conclusions drawn from them are misleading.

Jackson washed his *Gemma* samples through a 1-mm sieve. According to Sullivan (*3*), Sellmer (*4*), and our own work, newly released *Gemma* range in