methods of procedure are specified in advance. Everitt correctly points out that the pursuit of statistical significance is unfortunate and damaging, but to recommend informality as an alternative is to invite a return to the not-so-long-ago days when psychiatric research had the deserved reputation for producing junk.

Unlike Everitt, I would support the statistician who couldn't or wouldn't help the psychiatrist with a 500-item questionnaire that had been administered to 100 depressed patients. The statistician, if he or she had several years of experience working in the mental disorders, probably knew better than the psychiatrist, who may have been new to psychiatric research, that there wasn't much left to learn about the dimensions underlying depression, that hundreds of factor analyses of rating scales applied to depressives had already been performed, and that virtually nothing of value would be gained by the performance of yet another such factor analysis. Knowledge in psychiatry, and the psychiatrist's career in research, would both have been better served by the specification and testing of hypotheses, perhaps by a *confirmatory* factor analysis (Everitt and Dunn, 1983).

The opinion implicit in the preceding paragraph is that a statistician who's had extensive experience in a medical or scientific specialty may sometimes have as much or even more knowledge than a person formally trained in that specialty. Does Everitt subscribe to such heresy? How would he recommend a statistician to act if there were a serious disagreement on substantive matters between the statistician and the subject matter "expert?"

ADDITIONAL REFERENCES

AMERICAN PSYCHIATRIC ASSOCIATION COMMITTEE ON NOMENCLATURE AND STATISTICS. (1980). *Diagnostic and Statistical Manual of Mental Disorders*, 3rd ed. American Psychiatric Association, Washington.

COHEN, J. (1960). A coefficient of agreement for nominal scales. *Ed. Psychol. Meas.* **20** 37–46.

EVERITT, B. S. and DUNN, G. (1983). *Advanced Methods of Data Exploration and Modelling*. Heinemann, London.

FLEISS, J. L., COHEN, J. and EVERITT, B. S. (1969). Large sample standard errors of kappa and weighted kappa. *Psychol. Bull.* **72** 323–327.

FLEISS, J. L. and ZUBIN, J. (1969). On the methods and theory of clustering. *Multivariate Behav. Res.* **4** 235–250.

KORAN, L. M. (1975). The reliability of clinical methods, data and judgments. *New England J. Med.* **293** 642–646 and 695–701.

SCHMIDT, H. O. and FONDA, C. P. (1956). The reliability of psychiatric diagnosis. *J. Abnorm. Soc. Psychol.* **52** 262–267.

SPITZER, R. L., COHEN, J., FLEISS, J. L. and ENDICOTT, J. (1967). Quantification of agreement in psychiatric diagnosis. *Arch. Gen. Psychiat.* **17** 83–87.

Comment: The Biometric Approach to Psychiatry

Joseph Zubin

Everitt points out that psychiatry for the last several decades has been trying to emerge from its phenomenological descriptive cocoon into a more objective science. Galton was not alone in demanding measurement and numbers as a sine qua non for attaining "the dignity of a science." Thorndike is quoted as saying that whatever exists, exists in some amount and therefore could be eventually subjected to measurement and counting. Lord Kelvin is quoted as saying that one cannot understand a phenomenon until it is subjected to measurement. Both Emil Kraepelin and Karl Jaspers were appreciative of the importance of objective data and their evaluation. Kraepelin (1896) indicated his interest in measurement in the following statement:

> "As soon as our methodology has sufficiently proved itself through experience with healthy individuals, it would be possible to approach the actual ultimate goal of these efforts, the investigation of the sick personality, especially of the inborn pathological disposition. . . . We, therefore, have first of all to investigate whether it is possible by means of psychological tests to determine individual deviations, which cannot be recognized by ordinary observation. If that succeeds, we would be in the position, through the quantitative determinations at our disposal, to establish the

Joseph Zubin is Research Career Scientist and Coordinator for Research and Development, Veterans Administration Medical Center, and Distinguished Research Professor of Psychiatry, University of Pittsburgh Medical School. His mailing address is Biometrics Research, Veteran's Administration Medical Center, Highland Drive, Pittsburgh, Pennsylvania 15206.

Institute of Mathematical Statistics is collaborating with JSTOR to digitize, preserve, and extend access to *Statistical Science*. www.jstor.org

> borderline between health and disease much more precisely and more validly than has been possible so far."

Jaspers (1963, page 667) points out that:

> "The biometric methods give us more than figures and correlations. They foster clarity in all fields in which biometric variations can be established. Moreover, through the application of these methods we have concrete experiences which we would never have had without them. . . ."

That science depends on measurement and the evaluation of statistical data as well as on its theoretical concepts is today a firm belief of all scientists, but it was not always so. Price (1975) points out that in the pre-Alexandrian period in ancient Greece, measurement and numbers were underdeveloped while conceptual frameworks were supreme. On the other hand, in ancient Babylonia measurement and numbers were highly developed but conceptual frameworks were nowhere to be found to encompass the measurements. Thus, the ancient Greeks could describe the movements of the planets geometrically but could not measure their position, while the Babylonians could pinpoint the location of the planets against the stationary stars, and predict their future positions but had no conception of the geometry of the heavens. Alexander the Great, by conquering both nations, brought their skills together and thus science emerged as a combination of conceptual theories which had to be validated by measurement. The state of psychiatry until recently was quite similar to the state of pre-Alexandrian science. On the one hand, statisticians like Ben Malzberg measured accurately first admission rates, discharge rates, mortality, etc., but had no inkling of what these meant, while the psychiatrists spun conceptual models but had no use for statistics. It was not until Jellinek (1937), the first biometrician at the Worcester State Hospital, came along that theory and measurement were yoked at least in the United States. It is interesting to note that Everitt himself is not a member of a statistical unit but of a biometric unit. It is clear that in order to be a useful statistician in psychiatry one must not only be a master of statistics but must also have at least an acquaintance if not a deep understanding of psychiatry and its concepts and theories. Otherwise data and theory pass each other by.

I was a student in psychology and psychometrics before I entered the Psychiatric Institute at Columbia University and was quite innocent of psychiatry, yet I felt that my training could be applicable to psychiatric problems and took as my model, E. M. Jellinek, with whom I became acquainted in 1936 after visiting his hospital as a member of the Mental Hospital Survey Team under the auspices of the National Committee for Mental Hygiene, The Rockfeller Foundation, and the United States Public Health Service. No research could be undertaken in his hospital unless it had his biometric approval and this procedure led to a most productive research output. I set my goal along these lines.

Statistics, like many other fields must be responsive to the Zeitgeist and its needs. Some of the earliest advances in statistics came in response to gambling needs and later they came in response to the eugenic movement under Galton and Pearson. The first clinical trial in psychiatry came in response to the focal infection theory as a cause of mental disorder (Cotton, 1922). Kopeloff and Cheney (1922) and Kopeloff and Kirby (1923) undertook to test this hypothesis by removing surgically all the foci of infection from one group of patients and comparing their outcome with an untreated control group. The only difference they could find was a higher mortality for the experimental group but no difference in outcome. They, however, merely reported the number and percentages of the improved cases but made no statistical test of the difference. I was the first to compute the statistical significance of the difference in percentages and found it to be nonsignificant (Zubin and Zubin, 1977).

Descriptive statistics continued to be the primary contribution to psychiatry until the advent of the drug era except for a few epidemiological studies comparing whites and blacks and socioeconomic strata. One of the early symposia on prevalence of mental disorder was conducted by the American Psychopathological Association in 1945 (Zubin, 1945).

With the advent of the somatic treatment era in the forties and the drug treatment era in the fifties a demand grew for the use of statistics and measurement in comparing outcome under treatment with control groups. Up until then, the changes in the behavior of patients were so minimal that there was no need for measurement. With the advent of the drug era, a great need arose to determine the efficacy of the drugs.

The speed of the behavioral changes were so dramatic that it challenged the therapist to provide measures of the degree of change and the corresponding relation to dosage. For the first time in psychiatric history, psychotic behavior could be altered before your very eyes and the need for measurement and evaluation became manifestly apparent. This caught the measurement experts—mostly psychologists—unprepared, since all the available clincial tests and tools provided only static measurements like intelligence tests or projective techniques which were geared to trait rather than state measurement. Rating scales had to be produced practically overnight by such pioneers as Lorr, Wittenborn, and Malamud. There is probably an apocryphal story told about the physiologist, Hoagland, who found some biochemical change in Malamud's patients and asked to see Malamud's data on the same patients. Malamud handed him his voluminous case histories. "Are there no numbers here on their behavior," Hoagland inquired, to which Malamud cried, "No, but if you want numbers I'll make them up for you." That night with the help of Sands he converted his descriptions of behavior into rating scales.

The success of the drug therapies also required better classification of patients in order to obtain more homogeneous groups. To fill this need, the free floating clinical interview had to be converted from a blunderbuss into a sharpshooting rifle. Systematic semistructured and structured interviews resulted. Their use in the US-UK Diagnostic Project (Cooper et al., 1972) and in the WHO International Pilot Study of Schizophrenia (1974) were the proving grounds in which these instruments were tested. These projects roused diagnosis from its academic lethargy and finally led to the developments of such instruments as the SADS, DIS, Research Diagnostic Criteria, and the DSM-III classification system.

Thus far we have described how statistics influenced diagnoses and treatment in psychopathology. Did psychopathology in turn influence statistics?

One of the fundamental problems in psychopathology as in other sciences is the problem of classification. Most of the statistical techniques utilized in classification were originally based on approaches borrowed from psychometrics. These techniques were based on establishing differences between groups by such means as item analyses, reliability measures, and validity measures. They were essentially based on the search for individual differences. Individual similarities which clinicians often searched for, were never the focus. I had been intrigued by this gap in our approach and provided one of the first techniques for discovering individual similarities through the provision of a clustering technique for grouping like-minded individuals (Zubin, 1938). Since then more powerful clustering techniques have been developed and often used effectively in the search for subgroups in diagnostic categories.

Unfortunately, as Everitt has pointed out, clustering techniques have proliferated but because of their blind application have not always yielded results which could be replicated. As Fleiss and Zubin (1969) have pointed out, unless a mathematical model for the expected results of clustering is provided, the results cannot be evaluated because there is no goal to compare them to. On the other hand, the blind application of factor analysis runs up against Everitt's critique as well as against the frequent failure of the data to meet the underlying assumptions of factor analysis such as linearity, continuity, and homoscedasticity. Consequently, the factor analytic results may be quite misleading.

Another contribution which psychopathology demanded from statistics was a better way of measuring agreement between raters than was afforded by the percentage of agreements. Under the pressure of this demand, κ was born which discounted chance agreement and base line produced similarities or discrepancies.

Another problem which psychopathologists have called attention to is the extension of the survival curve analysis and hazard analysis from a single trial to determine the fate of a patient on comparative drug treatments, to a consideration of determining the outcome of illness for the patient himself over a stated follow-up period. This would include continuous monitoring of the patient's status over a specified period including the number of days he was experiencing an episode, and the number of relapses or new episodes he underwent. Some time ago Burdock and Hardesty (1961) proposed such an outcome index. Crude as that index was—it seemed to be related to the subjective clinical impression of outcome—it was empirically geared to a specific situation. There is need for developing an index for more general use.

Some of the other outstanding problems requiring attention are: (1) the development of better measures of similarity between individual profiles on categorical data and between individual curves on continuous data, (2) safeguarding against bias, (3) the problem of competing risks in follow-up studies, and (4) categorical versus dimensional approaches to classification.

The problem of classifying individual categorical profiles according to their similarity and the corresponding problem of classifying continuous individual tracings like EEGs is one which becomes especially important in genetic and familial investigations where a quantitative measure of degree of similarity is needed. At the present time, despite the helpfulness of extant techniques, there seems to be no better method than eyeballing. Why computerized methods for such comparisons have not been developed is somewhat of a mystery to me. Perhaps there are some techniques already available but they have not yet reached the level of common knowledge.

If one were to select the most grievous difficulty facing research in psychopathology, the problem of bias would be foremost. It has dogged research in psychopathology from the very beginning. The problem arises from the lack of specific definitions of the classification system. Much of the earlier research on schizophrenia suffered from this difficulty and has rendered the results of the research of that era doubtful. It is also likely that many of our conclusions regarding the course of schizophrenia and its long term outcome are to be faulted on the bias in the selection of cases for study. In fact, Manfred Bleuler (1972), the son of Eugen Bleuler (1911), attributes his father's belief in the chronicity of schizophrenia to the types of patients that perennially filled the clinics. These consisted primarily of the poorer outcome patients, and therefore, reinforced the feeling that schizophrenia was chronic. This bias may have been the reason why Kraepelin and his colleagues regarded schizophrenia as a deteriorating disorder. That this selective bias plagues follow-up studies in other fields also is amply demonstrated by Cohen and Cohen (1984). Manfred Bleuler (1972), who followed a cohort from their early age, came to the opposite conclusions regarding chronicity. He found that most patients, even some of long standing illness, improve. How to avoid this bias is one of the primary concerns of experimental designers. Even randomization which will remove the bias in the selection from a given universe will not relieve the problem of the bias in the selection factors in the universe itself. Only sampling from a specified general population whose demographic and ecological characteristics are known can serve to reduce or control the bias, but for rare disorders, sampling from the general population would turn out to be too expensive.

The World Health Organization has attempted to resolve this problem by conducting cross-cultural studies on incidence (first contact) of schizophrenia. However, because of difficulties in studying general populations, the investigation was based on individuals who sought help for the first time for whatever mental problems they presented. The mentally ill with schizophrenia were screened out for investigation. That this method will suffer from cultural bias toward recognition and detection of mental illness is obvious. But ignoring this bias, Häfner (1986) proceeded to present data showing that if the diagnosis on this (biased) sample is limited to the category of nuclear schizophrenia the rates of incidence of schizophrenia were found to be nondifferential across the different cultures. Häfner proceeded to interpret these findings as indicating that sociocultural and other ecological factors did not seem to have any influence on the etiology of schizophrenia. In addition to the acknowledged source of bias, there is still another factor that casts doubt on his conclusion. Even if no differential in incidence rates was found in the investigation of a general population freed from the bias described above, the results would still be suspect. Since general population surveys deal with the phenotype and not with the genotype, there is good reason to believe that culture may still get its licks in determining incidence. If we assume that populations vary in the threshold for perception and recognition of deviance in behavior, it is quite likely that populations with basically high incidence will tend to recognize the very severe cases and fail to recognize the mild deviants. Only when markers of vulnerability independent of the phenotype become available can the question be resolved as to whether culture influences etiology.

Other examples of biases which distorted the results of investigations in psychopathology are not difficult to find. Kallmann's inflation of the degree of concordance in monozygotic twins (Rosenthal, 1962) is a classic example of bias in selection. The bias in the clinical opinion that Buerger's disease was limited to Jews was discovered when the King of England was found to be a victim of this disorder. In contrast with this type of clinical fallacy there is the ecological fallacy of attributing to the correlation between two traits within individuals the correlation found between these traits for the means of groups to which the individual belongs. In this way the correlation between average rental and rate of schizophrenia by city areas in Chicago (Faris and Dunham, 1939) was thought to hold true for individuals until the Bristol study (Hare, 1956) indicated that the correlation was due large to the proportion of loners living in the city areas.

Another fallacy is to regard negative family history as an indicator of sporadic schizophrenia without considering the problem of penetrance and making sure that the number of first degree relatives is sufficiently large to rule out the absence of schizophrenia in the family. Some preliminary estimates indicate that n must exceed 21 for a penetrance of .26.

That competing risks such as mortality can produce problems in the evaluation of outcome of treatment or of illness has been studied for a long time. In one ECT study inclusion or exclusion of mortality was a factor in determining the evaluation until Neyman's technique for competing risks was applied (Neyman, 1950).

Some of the statistical problems facing the psychostatistician (a statistician concerned with mental rather than somatic disorders) are universal problems. Thus, the problem of categorical versus dimensional analyses has been a perennial problem. It may depend on the state of the art rather than on logic (Zubin, 1955). These is a need for developing a set of principles for converting dimensional approaches to categorical approaches even as theorems about points can be converted to theorems about lines in projective geometry. Here, again, perhaps such methods are already available but they have not yet reached the consumer.

All in all, there seems to be a great need for a middle man between the new developments in statistical knowledge and the final consumer.

These are some of the issues which come to mind in reading Everitt's thoughtful comments on the current status of statistics in psychiatry.

ADDITIONAL REFERENCES

BLEULER, E. (1911). *Dementia Praecox or the Group of Schizophrenias.* International University Press, New York. (Translated by J. Zinkin, 1950.)

BLEULER, M. (1972). *Die Schizophrenen Geistesstörungen im Lichte langjähriger Kranken und Familiengeschichten.* Intercontinental Medical Book Corp., New York.

BURDOCK, E. I. and HARDESTY, A. S. (1961). An outcome index for mental hospital patients. *J. Abnorm. Soc. Psychol.* **63** 666–670.

COHEN, P. and COHEN, J. (1984). The clinicians' illusions. *Arch. Gen. Psychiatry* **41** 1178–1182.

COOPER, J. E., KENDELL, R. E., GURLAND, B. J., SHARPE, L., COPELAND, J. R. M. and SIMON, R. (1972). *Psychiatric Diagnosis in New York and London: A Comparative Study of Mental Health Admissions.* Maudsley Monograph No. 20. Oxford Univ. Press, London.

COTTON, H. A. (1922). The etiology and treatment of the so-called functional psychoses. *Amer. J. Psychiatry* **2** 157–210.

FARIS, R. E. L. and DUNHAM, H. W. (1939). *Mental Disorders in Urban Areas: An Ecological Study of Schizophrenia and Other Psychoses.* Univ. Chicago Press, Chicago.

FLEISS, J. L. and ZUBIN, J. (1969). On the methods and theory of clustering. *Multivariate Behav. Res.* **4** 235–250.

HÄFNER, H. (1986). Epidemiology of schizophrenia. Paper presented at Session on Epidemiology and Course of Schizophrenia, in Symposium on Search for the Causes of Schizophrenia, at the Occasion of the 600th Anniversary of the University of Heidelberg, Germany.

HARE, E. H. (1956). Mental illness and social conditions in Bristol. *J. Ment. Sci.* **102** 349–357.

JASPERS, K., ed. (1963). *General Psychopathology.* Translated by J. Hoenig and M. W. Hamilton. Univ. of Chicago Press, Chicago.

JELLINEK, E. M., ed. (1937). *Biometric Bulletin.*

KOPELOFF, N. and CHENEY, C. O. (1922). Studies in focal infection: Its presence and elimination in the functional psychoses. *Amer. J. Psychiatry* **2** 139–156.

KOPELOFF, N. and KIRBY, H. G. (1923). Focal infection and mental disease. *Amer. J. Psychiatry* **3** 149–198.

KRAEPELIN, E. (1896). Der psychologische Versuch in der Psychiatrie. *Psychologische Arbeiten* **1** 77.

NEYMAN, J. (1950). *First Course in Probability and Statistics.* Holt, Rinehart and Winston, New York.

PRICE, D. DE S. (1975). The peculiarity of the scientific civilization. In *Science Since Babylon*, enlarged ed. Yale Univ. Press, New Haven, Conn.

ROSENTHAL, D. (1962). Problems of sampling and diagnosis in the major twin studies of schizophrenia. *J. Psychiatric Res.* **1** 116–134.

SARTORIUS, N., JABLENSKY, A., KORTEN, A., ERNBERG, G., ANKER, M., COOPER, J. E. and DAY, R. (1986). Early manifestations and first-contact incidence of schizophrenia in different cultures. *Psychol. Med.* **16** 909–928.

WORLD HEALTH ORGANIZATION. (1974). *International Pilot Study of Schizophrenia.* Wiley, Geneva.

ZUBIN, D. and ZUBIN, J. (1977). From speculation to empiricism in the study of mental disorder: Research at the New York State Psychiatric Institute in the first half of the 20th century. In *Roots of American Psychology: Historical Influences and Implications for the Future* (R. W. Rieber and K. Salzinger, eds.) 104–135. Annals of the New York Academy of Sciences, New York.

ZUBIN, J. (1938). Sociobiological types and methods for their isolation. *Psychiatry* **1** 237–247.

ZUBIN, J., ed. (1945). *Trends of Mental Disease.* Kings Crown Press, New York.

ZUBIN, J. (1955). Clinical versus actuarial prediction: a pseudoproblem. *Proc. of the 1955 Invitational Conference on Testing Problems* 107–128. Educational Testing Service, Princeton, N. J.

Comment: Psychiatric Statistics and Clinical Information

Juan E. Mezzich and Chul Woo Ahn

The role of statistics in psychiatry is broad and emerging, attributes also applicable to psychiatry itself, a clinical science with biological and psychosocial underpinnings. It may be helpful to give attention to historical perspectives and informational complexity in order to understand the current role of statistics in psychiatry as well as to be prepared to appraise its future directions.

Juan E. Mezzich is Professor of Psychiatry and Director, Clinical Information Systems; Chul Woo Ahn is Research Statistician, Clinical Information Systems, Western Psychiatric Institute and Clinic, University of Pittsburgh, 3811 O'Hara Street, Pittsburgh, Pennsylvania 15213.

Early endeavors to use quantitative methods in clinical and epidemiological psychiatry are best illustrated by the pioneering works of Philippe Pinel, Jean Etienne Esquirol, and J. B. M. Parchappe in the late seventeenth century and first half of the eighteenth century. Their guiding principles were careful observations of clinical events and a critical and quantitative investigative approach. These French pathfinders are responsible for what is probably the earliest documented use of statistics in psychopathology and psychiatric care. Of particular interest here is Parchappe (1839) who, in his *Recherches Statistiques sur les Causes de l'Aliénation Mentale*, not only presented frequency analyses of patients and patient-related events, but also ascertained relationships between complex domains, i.e., physical and moral causes and