Note

"Three approaches to resolving problems arising from assumption violation during statistical analysis in geographical research"

Bruce Mitchell

Cahiers de géographie du Québec, vol. 18, n° 45, 1974, p. 507-523.

CAHIERS DE GÉOGRAPHIE DE QUÉBEC,
Vol. 18, No. 45, décembre 1974, 507-523

THREE APPROACHES TO RESOLVING PROBLEMS ARISING FROM ASSUMPTION VIOLATION DURING STATISTICAL ANALYSIS IN GEOGRAPHICAL RESEARCH

At least two fundamental concerns arise when using statistical analysis in geographical research. One focuses upon whether or not the data satisfy the assumptions associated with given tests. The second arises from concern that data satisfying required assumptions are at a tangent to the interests of geographers [1]. While both points deserve thought and attention, this paper focuses upon the former within the context of the Neyman-Pearson approach to inferential statistical analysis [2].

Given such an orientation, three inter-related considerations appear. First, it is essential to identify the assumptions which, if not satisfied, will lead to difficulties in interpreting test results. Second, the researcher should develop some personally-satisfying criteria which may be used to decide whether assumptions are or not satisfied. Third, if an assumption is violated by a data set, the investigator must decide how to resolve this problem. Minor attention is directed to the first two considerations noted here. The main thrust of the paper is toward identifying methods for proceeding when assumptions are not satisfied.

ASSUMPTION IDENTIFICATION AND CRITERIA

Independence of observations, normality and homogeneity frequently are assumed to underlay data during statistical analysis within the Neyman-Pearson school. A further assumption often encountered is that of linearity. Thus, as a basic guideline, the geographer often has to consider whether data meet one, all, or some combination of these four assumptions. Certainly the requirement of independence will always be encountered. The relevance of the other three is a function of the specific tests used in analysis.

Rare is the analyst who possesses data which perfectly meet the assumptions of normality, homogeneity or linearity. Minor deviations from these ideals are the norm, and major deviations are not uncommon [3]. The

---

1 GOULD, P. (1970) Is « Statistix Inferens » the geographical name for a wild goose ? *Economic Geography,* 46 : 443.

2 MITCHELL, B. and J. MITCHELL. (1973) *Inferential statistical analysis : issues, foundations and schools of thought.* Council of Planning Librarians Exchange Bibliography No. 372. Monticello, Council of Planning Librarians : 17-19.

3 BONEAU, C. A. (1960) The effects of violations of assumptions underlying the t test. *Psychological Bulletin,* 57 : 49.

difficulty is, of course, to differentiate between a minor or acceptable deviation and a major one. The outcome is the need for criteria against which the investigator judges the data. Development and presentation of explicit criteria would assist in eliminating misunderstandings between researchers. Use of such criteria would also be necessary, but not sufficient, to encourage validation and replication of research findings.

General guidelines are difficult, and perhaps necessarily impossible, to develop. The nature of the problem under investigation, the consequences of faulty interpretations, the availability of external checks and the experience of the investigator all will and should result in different judgements being made about the adequacy of data. It thus becomes a task for each researcher to develop a set of personally-satisfying criteria which are made available to readers.

PROCEDURES WHEN ASSUMPTIONS ARE VIOLATED

After having identified critical assumptions and having determined whether or not they are satisfied, the investigator has a final task. In what manner should analysis proceed given that one or more assumptions are not met ? Three alternatives are considered here with the qualification that the ultimate goal is to select the most powerful and efficient test available. With this constraint, the researcher may either turn to transformations, distribution-free or nonparametric tests, or the concept of robustness. The first two alternatives, transformations and distribution-free or nonparametric methods, are discussed briefly. More time is directed toward robustness for several reasons. Relatively few geographers appear to be aware of this concept. Or, if aware of robustness, they do not use it explicitly in published research. Conversely, where geographers have referred to robustness they often have misused or abused the concept.

TRANSFORMATIONS

Bartlett notes that when the original distribution is « grossly non-normal », it is often possible by a suitable transformation to obtain a distribution more nearly normal [4]. Boneau agrees with such a statement, as he notes that « . . data have an exasperating tendency to manifest themselves in a form which violates one or more assumptions underlying the usual statistical tests of significance. » [5] Encountering such data, Boneau observes that the researcher usually attempts transformations in order to make the assumptions tenable, or looks elsewhere for alternative tests.

What is a transformation, and what are the general situations for which they are applicable ? Mueller provides useful answers to both questions. In

---

[4] BARTLETT, M.S. (1935) The effect of non-normality on the t-distribution. *Proceedings of the Cambridge Philosophical Society*, 31 : 223.

[5] BONEAU, *Op. cit.* 49.

his view, the term transformation refers to operations by which one set of numbers is changed into another set. [6] Examples are the familiar procedures of computing logarithms or reciprocals of observations.

Mueller suggests that transformations are used most frequently in three situations.[7] The first situation is one in which a quantitative theoretical statement or prediction is reduced to a simpler form in order to facilitate inspection of the agreement between data and theory. To illustrate, he observes that reduction of a theoretical statement to a simple form often begins with an effort to express the relationship in linear form. A second situation is one in which a transformation is used to aid description, such as to find an approximate descriptive expression for the relationship between variables. As opposed to the first situation already described, such a quantitative expression generally has no theoretical basis. Instead, the expression simply describes the data within constraints established by the consistency of the data and concern for simplicity. The important aspect to stress is that the observed relationship in the transformed data implies no theory and the resultant descriptive expression does not constitute a theory of the data under investigation.

The third situation occurs when one frequency distribution is changed to another to permit application of more efficient statistical tests during analysis. Mueller states that although conversions of observations into any specified frequency function is possible, the most frequently encountered problem is that of making data approximate the normal curve. In his words, the reason « . . . arises mainly from the greater amount of information available on the sampling characteristics of the parameters of this function ; if the data can be put into the appropriate form, larger resources exist for testing statistical hypotheses. » [8]

A useful framework is thus established against which the geographer can consider the role of transformations when analyzing data. Mueller makes an important point that the investigator must distinguish between theoretical and computational implications of transforming data. Transformations may or may not have a theoretical underpinning, and this aspect must be kept in mind when interpreting the test statistics stemming from such data.

The issue of theoretical versus operational rationals for transforming data provides a departure point for considering some difficulties which may confound interpretation of test results following transformations. Two aspects are important. First, transformations may imply a relationship which is not present in the raw data. As Gould has noted with reference to regression or correlation analysis, transformation of variables to satisfy assumptions may

---

[6] MUELLER, C. G. (1949) Numerical transformations in the analysis of experimental data. *Psychological Bulletin*, 46 : 198.

[7] *Ibid.* 198-208.

[8] *Ibid.* 208.

change the very form of the relationship and model.[9] To illustrate, he notes that transforming both variables to logarithms implies the existence of an exponential relationship between the variables. Another complication may arise if, by transforming to satisfy one assumption such as normality or linearity, another assumption is violated. It has been pointed out, for example, that in testing a theoretical prediction of linearity of relationship between two transformed variables, the transforming operations may introduce non-normality and heterogeneity of variability around the theoretical line.[10]

In discussing the analysis of variance F-distribution, Lindquist elaborates upon the second consideration.[11] He concludes that if the treatments cause the variances of observations to differ but do not create differences among means, or if the treatments cause variances to differ independently of the differences among the means, then no valid transformation is available. In addition, he notes that if the distributions have the same variances but differ in shape of distribution, no valid transformation is possible. The reason for these caveats, of course, is that transformations attempting to normalize different distributions with equal variances may generate acceptable distributions but heterogeneous variances. Thus, in attempting to resolve one problem through transformations, one or more new ones may be created.

The conclusions to be drawn from this discussion are straight forward. Transformations constitute a useful approach for overcoming violation of assumptions during inferential statistical analysis. Transformations may be utilized when it is necessary to achieve linearity, normality or homogeneity. Conversely, transformations are not without problems. One assumption may be realized at the expense of generating a new violation of another assumption. The basic relationship or model underlaying the data or analytical method may not be consistent with data in their transformed state. Furthermore, no theoretical foundation may support the transformation. Yet the geographer's concern with pattern and process may tempt inferences about these dimensions based upon data which are difficult if not impossible to interpret due to distortions created by transformations. Gould has stated succinctly the predicament that transformations may create when he exclaimed that, « Too often we end up relating the value of one variable to the log of another, with the square root of the third, the arc sin of the fourth, and the log of a log of a fifth. Everything is normal, statistically significant at the one percent level — except that we have not the faintest idea what it means. »[12] Given that transformations are a useful yet imperfect approach, the geographer should be aware of other alternatives which may be pursued.

---

9 GOULD. *Op. cit.* 442.

10 MUELLER. *Op. cit.* 220-221.

11 LINDQUIST, E. F. (1953) *Design and analysis of experiments in psychology and education.* Boston, Houghton Mifflin Co.

12 GOULD, *Op. cit.* 442.

DISTRIBUTION — FREE OR NONPARAMETRIC TECHNIQUES

Most investigators would agree that the « best » inferential statistical test is that which is most powerful. While power is an essential criterion, those tests enjoying the highest power generally tend to have the greatest number of assumptions or constraints. Thus, parametric techniques, usually having high power, are also characterized by such assumptions as independence of observations, normality and homogeneity. An alternative is to utilize distribution-free or nonparametric tests [13] which have fewer assumptions but are generally considered to be less powerful. Consequently, these less restrictive techniques are frequently viewed as less desirable since they are not as likely as their parametric alternatives to catch a significant difference.

Numerous writers attest to the alleged weakness of distribution-free or nonparametric tests. French, writing for geographers, has stated that a « . . . non-parametric test, because of its much weaker assumptions than a parametric test . . . is generally a less powerful test than a parametric one. » [14] He continues, however, to argue that the power of nonparametric methods may be improved by increasing the size of the sample relative to that of a comparable parametric test. This procedure is generally a valid one. Unfortunately, it negates his earlier argument that nonparametric tests have great merit in the small sample situation. The reason cited by French is that with small samples the form of the distribution from which the sample is taken is not sufficiently well known for it to be regarded as normal. French's argument creates a paradox. On the one hand, nonparametric tests are useful in small sample cases for which the shape of the distribution is indeterminant. On the other hand, increased power is attained for nonparametric tests by increasing sample size. This argument by French, by itself, does not provide a strong recommendation for nonparametric methods.

Others have written in a manner which supports French. Cole and King have suggested that inferential tests may essentially be divided into two types:

---

[13] BRADLEY, J.V. (1968) *Distribution-free statistical tests.* Englewood Cliffs, Prentice-Hall ; CONOVER, W. J. (1971) *Practical nonparametric statistics.* New York, John Wiley ; EDGINTON, E. S. (1969) *Statistical inference : the distribution-free approach.* New York, McGraw Hill ; FERGUSON, G. A. (1965) *Nonparametric trend analysis.* Montréal, McGill University Press ; FRASER, D. E. S. (1957) *Nonparametric methods in statistics.* New York, John Wiley ; GIBBONS, J. D. (1971) *Nonparametric statistical inference.* New York, McGraw Hill ; HAJEK, J. (1969) *A course in nonparametric methods.* San Francisco, Holden-Day ; KRAFT, C. H. and C. VAN EEDEN (1968) *A nonparametric introduction to statistics.* New York, Macmillan ; NOETHER, G. E. (1967) *Elements of nonparametric statistics.* New York, John Wiley ; PURI, M. L. ed. (1970) Nonparametric techniques in statistical inference. Cambridge, Cambridge University Press ; SIEGEL, S. (1956) *Nonparametric statistics.* New York, McGraw Hill ; WALSH, J. E. (1962) *Handbook of nonparametric statistics.* Princeton, D. Van Nostrand, 3 volumes.

[14] FRENCH, H. M. (1971) Quantitative methods and non-parametric statistics. *Quantitative and qualitative geography.* H. M. French and J. B. Racine eds., Ottawa, University of Ottawa, Department of Geography, Occasionel Paper No. 1 : 123.

nonparametric and parametric.[15] With regard to nonparametric methods, they conclude that « . . . . on the whole these tend to be less powerful than the parametric tests ». Writers from other disciplines reinforce this viewpoint. Lindquist, a psychologist, has commented that «. . . all of these (nonparametric) tests are less powerful than those assuming normality and homogeneity of variance. »[16] Another psychologist has also referred to the lack of power of nonparametric tests as being a decided handicap in stimulating research.[17] Thus, a viewpoint has been perpetuated under which nonparametric tests appear as less powerful methods relative to their parametric counterparts. The remaining discussion in this section attempts to demonstrate that this viewpoint is misguided, and has led to a distorted impression of the nature and comparative values of parametric and nonparametric tests of significance.

Before discussing the fallacy associated with the relative power of nonparametric tests, it is necessary to define these tests and also to outline the manner in which power is being used. Conover has stated that « There is no agreement among statisticians as to the meaning of the word nonparametric. In fact there is not even agreement among statisticians concerning whether certain tests should be classified as parametric or nonparametric. »[18]

With this appropriate caution about the status of nonparametric techniques, this paper adopts Bradley's definition.[19] He observes that although the terms distribution-free and nonparametric are not synonymous, popular usage has equated them. In his view, a nonparametric test makes no hypothesis about the value of a parameter in a statistical density function. Distribution-free tests, on the other hand, make no assumptions about the precise form of the sampled distribution. These definitions are not mutually exclusive, and one test may have the characteristics of both distribution-free and nonparametric tests. Nevertheless, it is the distribution-free term which comes closest to describing the quality which makes such tests attractive to geographers.

Power and power-efficiency are two further terms requiring definition. Power may be defined as the probability of rejecting the null hypothesis when it is in fact false. The concept of power-efficiency relates to the amount of increase in sample size necessary to make one test as powerful as another. In order to determine the power efficiency of a test, the usual procedure is to determine the number of observations needed by the first test to obtain the same power as the second test with a specified number of cases. As a result, power efficiency is generally used to indicate the power of a given test relative to its most powerful alternative. For example, the Mann-Whitney test would be compared to the Student test.

---

15 COLE, J. P. and C. A. M. KING (1968) *Quantitative geography.* London, John Wiley.

16 LINDQUIST. *Op. cit.* 90.

17 BONEAU. *Op. cit.* 49.

18 CONOVER. *Op. cit.* 93.

19 BRADLEY. *Op. cit.* 15.

Blalock elaborates on the nature of power when he states that « If we refer to the power efficiency of a nonparametric test as 95%, we mean that the power of the nonparametric test using 100 cases is the same as that of the t test using 95 cases if the model used in the t test is correct. » [20] Continuing, Blalock notes that it is necessary to assume a given form for the population in order to make the evaluation of power. As a result, in determining the power efficiency of a distribution-free or nonparametric test the researcher is asking how much a failure to accept a normality assumption will cost if in fact such an assumption were legitimate. In the example Blalock presents, the failure to accept a normality assumption and the subsequent use of a nonparametric test « costs » an extra five observations above the number needed in the t test.

The fallacy involved in calculating power efficiency relates to the procedure of comparing two tests when assuming that all the requirements of the more powerful test are satisfied. If all requirements of the more powerful one were satisfied, the investigator would be foolish to consider the nonparametric alternative. Bradley has succinctly outlined the nature of this fallacy. In his words

> . . . the earliest efficiency figures were obtained by comparing the distribution-free test with a parametric test under common conditions meeting all the assumptions of the latter. Thus the parametric test was permitted both to hurl the challenge and to choose the weapons. Under these loaded conditions the best parametric test was found to be more efficient than (or, at worst, equally efficient to) its distribution-free competitor. [21]

Bradley went on to observe that the actual difference in power-efficiency was often remarkably small. Unfortunately most attention was directed to the existence of the unfavorable direction of the difference rather than towards the fact that the absolute difference was frequently of small extent.

It is hoped that the preceding discussion will lay to rest the prevailing notion, as expressed in the quotes by French and Cole and King, that nonparametric tests are less powerful than their parametric counterparts. The fact such evaluations have assumed that the assumptions of the parametric tests were satisfied makes the value of comparisons questionable.

In addition, Blalock has noted that the power efficiency of a test is a function of numerous variables. [22] The sample size utilized is a basic element. A test may be highly efficient for large samples but much less efficient for smaller ones, or vice versa. The choice of significance level, and the use of simple or composite research hypotheses also influence power-efficiency. As a result, blanket statements that a given technique is less powerful than another can have little meaning unless qualified with a statement about the sample size, significance level and type of research hypothesis involved. An

---

[20] BLALOCK, H. M. (1960) *Social Statistics.* New York, McGraw Hill : 192.
[21] BRADLEY. *Op. cit.* 12.
[22] BLALOCK. *Op. cit.* 192-193.

> Compared with the most powerful parametric test, the F test, under conditions where the assumptions associated with the statistical model of the F test are met, the Kruskal-Wallis test has asympotic efficiency of 95.5. percent. [23]

illustration of the type of statements which have little meaning is the following Not only does this statement assume that the requirements of the parametric test are satisfied, it also provides no information about sample size, level of significance or type of test.

Like transformations, nonparametric or distribution-free tests do not offer a perfect alternative when assumptions can not be satisfied. They often can not accommodate multivariate research designs, do not incorporate higher order interactions and occasionally have an imperfectly defined rejection region. Despite these real weaknesses, however, the techniques are an alternative, and have been unfairly criticized for their lack of power.

ROBUSTNESS

To this point the discussion has taken for granted that violation of test assumptions is unacceptable. Faced with data which do not meet the requirements of inferential tests, the alternatives put forward necessitate modification of the data or reliance on tests with less restrictive constraints. A third alternative arises if it can be determined that violation of assumptions does not affect probability distributions and test results. In other words, the test is said to be « robust », or insensitive to violations of underlying requirements of the test model.

Statisticians have long been aware of, and concerned about, the difficulties created by assumptions originally made for mathematical convenience. As a result, considerable attention has been focused upon determining what impact violations of assumptions have upon test results. As early as 1929, for example, Pearson noted that « One of the most important problems with which the mathematical statistician is faced is that of bringing his theoretical structures into some degree of correspondence with the situations of practical experience. » [24] Particularly in the small sample case, Pearson believed two problems arise. First, the population may not be completely stable. Second, even if the researcher is sure of population stability, « . . . it will generally be impossible for him to obtain any certain estimate of its exact form. » [25]

Pearson then posed the type of questions which analysts frequently must raise. In his words, the question becomes

> I do not know whether my distribution is normal, although from my general experience in the past I do not think it is likely to be excessively skew

---

[23] SEIGEL. *Op. cit.* 192-193.

[24] PEARSON, E. S. *et. al.* (1929) The distribution of frequency constants in small samples from non-normal symmetrical and skew populations. *Biometrika,* 21 : 259.

[25] *Loc. cit.*

> or leptokurtic. How sensitive are the « normal theory..» tests to changes in population form ? May I use some with less hesitation than others ? [26]

In his study Pearson attempted to answer these questions with regard to the Student t distribution when dealing with the mean of a single sample and when considering the difference between the means of two samples. In both situations, after conducting experiments, he concluded that « . . . the differences between them (observed frequencies) and the « normal theory » values . . . are hardly large enough to lead to any serious errors in inference. » [27]

This introductory work by Pearson was pursued by numerous other statisticians. Bartlett confirmed that « . . . for moderate departures from normality this test (Student t) may still be used with confidence », although no clear definition of « moderate departures from normality » was presented. [28] Pearson [29], Geary [30], David and Johnson [31] and Lindquist [32] presented similar findings for the F-distribution when normality was not met. Walsh [33], Horsnell [34] and Box [35] showed that the F-distribution was not affected significantly when homogeneity was violated. Daniel [36] demonstrated that lack of independence did not alter the t distribution, and Boneau [37] showed that the t- and F-distributions seemed affected only marginally by violations of both normality and homogeneity.

A body of literature thus developed in which evidence suggested at least some parametric inferential tests were not affected by assumption violations. This information was noted by critics of nonparametric methods. [38] Such critics could argue convincingly that if parametric tests were not affected by violations of assumptions then they should be used instead of their less powerful nonparametric competitors.

---

[26] *Loc. cit.*

[27] *Ibid.* 274.

[28] BARTLETT. *Op. cit.* 231.

[29] PEARSON, E. S. (1931) The analysis of variance in cases of non-normal variation. *Biometrika,* 23 : 114-133.

[30] GEARY, R. C. (1936) The distribution of « student's » ratio for non-normal samples. *Supplement, Journal of the Royal Statistical Society,* 3 : 178-184 ; GEARY, R. C. (1947) Testing for normality. *Biometrika,* 34 : 209-242.

[31] DAVID, F. N. and N. L. JOHNSON (1951) The effect of non-normality on the power function of the F-test in the analysis of variance. *Biometrika,* 38 : 43-57.

[32] LINDQUIST. *Op cit.* 90.

[33] WELSH, B. L. (1937) The significance of the difference between two means when the population variances are unequal. *Biometrika,* 29 : 350-362.

[34] HORSNELL, G. (1953) The effect of unequal group variances on the F-test for the homogeneity of group means. *Biometrika,* 40 : 128-136.

[35] BOX, G. E. P. (1953) Non-normality and tests on variances. *Biometrika,* 40 : 318-335.

[36] DANIELS, H. E. (1938) The effect of departures from ideal conditions other than non-normality on the t and z tests of significance. *Proceedings of the Cambridge Philosophical Society,* 34 : 321-328.

[37] BONEAU. *Op. cit.*

[38] *Ibid.* 51 ; LINDQUIST. *Op. cit.* 86.

Two important characteristics marked this literature. First, the authors were not clear as to the extent of violations which resulted in no effect on test results. In fact, the impression was usually left that severe deviations would have an impact, even though that which constituted a « severe deviation » was not clarified. As Bartlett explained in his study,

> An important point to notice is that the form of derived distributions when the original distribution is grossly non-normal does not concern us, for the usual tests of significance will then not be the appropriate ones to use. In such cases it is in practice often possible by a suitable transformation to obtain a distribution more nearly normal. [39]

The second characteristic related to the rather specific conditions which would lead to insensitivity of assumption violation. An example is provided in Boneau's conclusions. [40]. While his final conclusion was that « the t test is seen to be functionally non-parametric or distribution-free », numerous qualifications were added. Thus, he stated that the t distribution was essentially impervious to violation of homogeneity and normality in the two sample situation if the two sample sizes were equal, or nearly equal, and the underlying populations, although not normal, had the same shape. If the distributions were skewed then the variances had to be equal. On the other hand, if the sample sizes were unequal, no difficulty would be encountered provided that the variances were compensatingly equal. However, a combination of unequal sample sizes and unequal variances « automatically produces inaccurate probability statements which can be quite different from the normal values ». On a more positive note, however, it could be argued that specific guidelines were being developed to help the investigator decide whether to proceed with statistical analysis when assumptions were violated.

The concept of robustness emerged from the empirical work cited above. Introduced by Box in 1953 [41] and then elaborated upon in 1955 [42], the term was used in the following sense. A test was robust against violation of a specified assumption if the probability of a Type I error was not changed. In the discussion which followed its introduction to the Royal Statistical Society, the eminent statistician Professor Barnard declared that strategic as opposed to detailed tactical advances in statistics were usually associated with the birth and christening of new ideas. [43] In his view, robustness was a new addition to the family of such useful concepts as « efficiency », « sufficiency », « likelihood » and « power », and it could be confidently predicted that « . . . the idea will have a long and vigorous life ».

---

39 BARTLETT. *Op. cit.* 223.

40 BONEAU. *Op. cit.* 61-62.

41 BOX. *Op. cit.*

42 BOX, G. E. P. and S. L. ANDERSEN (1955) Permutation theory in the derivation of robust criteria and the study of departures from assumption. *Journal of the Royal Statistical Society,* Series B, 17 : 1-34.

43 *Ibid.* 32.

In presenting robustness, Box made a number of important points. He argued that statistical tests should be sensitive to changes in specific factors under test, and insensitive to changes in extraneous factors not under test [44]. A test satisfying the first requirement met the notion of « power », while a test satisfying the second should be labelled as « robust ». In his view, parametric tests tend to satisfy the first requirement, at least when assumptions are true, but not necessarily the second. Conversely, non-parametric tests tend to satisfy the second requirement but not necessarily the first.

Box noted several interesting implications of robustness. [45] First, he observed that it was a usual practice to conduct a test of homogeneity of variances prior to making an analysis of variance test for equal means. He suggested, based on his research, that when little is known about the parent distribution, such a practice could lead to more wrong conclusions than if the test for variances were omitted. Evidence available showed the analysis of variance test was affected surprisingly little by variance inequalities when group sizes were equal. Since the analysis of variance test was also known to be insensitive to non-normality, he felt that the test could be used safely under most practical conditions. Thus, in his view,

> To make the preliminary test on variances is rather like putting to sea in a rowing boat to find out whether conditions are sufficiently calm for an ocean liner to leave port ! [46]

A further consequence resulting from his concept deserves mention. Box believed that robustness was even more important in that, for practical purposes, the analyst should seek to utilize a procedure in which the statistical test has maximum power and for which the statistic involved is fully efficient. [47] When trade-offs had to be made, he believed that the qualities of power and efficiency should be sacrificed to ensure the former quality. Given this viewpoint, however, he did not feel that « . . we need necessarily go the extreme of using non-parametric tests when it may well be that more powerful robust parametric tests can be found ».

With the concept of robustness available, it might seem that geographers have an attractive alternative available when data do not meet test assumptions. Rather than conducting laborious transformations, with the attendant interpretation problems, or turning to nonparametric tests with their constraints, the geographer could cite robustness as justification for proceeding when little was known about the nature of population characteristics or else when it was known assumptions were not satisfied. If this stance is a valid one, then little concern would be created by comments such as the one by French that « If these assumptions (normality, independence, homogeneity)

---

[44] *Ibid.* 1.
[45] BOX, (1953) *Op. cit.* 333.
[46] *Loc. cit.*
[47] *Loc. cit.*

are not met in full, . . ., it is impossible to estimate the validity and power of a parametric test. » [48] To determine whether robustness represents such an alternative, the following discussion considers a more careful definition of the term, its use by geographers, and comments on such use.

Concerning definitions, Box and Tiao distinguish between test robustness and inference robustness. [49] Test robustness, or the changes in a probability distribution when the parent distribution deviates from the form assumed, is the concern in this discussion. This form of robustness should be kept distinct from inference robustness, or changes in such aspects as the significance level when different tests are chosen to adjust to changes in the parent distribution.

Despite this attempt at clarification by Box, Huber has concluded that « From the beginning, « robustness » has been a rather vague concept. . . ». [50] In his view, if an analyst wishes to rationally choose between different tests on the basis of robustness, the goals being sought have to be made precise. From his review of the literature, Huber concludes that unfortunately « . . . a consensus has not been reached ; although the goals rarely are stated in an explicit fashion, one can discern at least five or six conflicting ones, and I do not think that all of them should be called by the same name, « robust ». [51] »

Huber describes five different ways in which robustness seems to be used. First, he believes some writers use the concept to signify a high absolute efficiency for *all* suitably smooth-shaped distributions. A second use is to denote a high efficiency *relative* to the sample mean or other selected estimates for all distributions. A third interpretation is for a high absolute efficiency over a strategically selected *finite* set of distribution shapes, such as the normal, logistic, double exponential, Cauchy, and rectangular. A derivative of the third viewpoint is a high absolute efficiency over a strategically selected family of distribution shapes. A fourth opinion sees robustness relating to a small asymptotic variance over some neighborhood of one shape, particularly that of the normal distribution. The fifth view considers that to be robust the distribution of the estimate should change little under arbitrary small variations of the underlying distribution. In addition, any variation should be uniform for a sample size « n ».

These varying definitions relate to the shape of the population distribution. Huber argues that in his opinion the fourth and fifth goals are the important ones for investigators. He views robustness as a « . . . kind of insurance problem : I am willing to pay a premium (a loss of efficiency of,

---

48 FRENCH. *Op. cit.* 121.

49 BOX, G. E. P. and G. C. TIAO (1964) A Bayesian approach to the importance of assumptions applied to the comparison of variances. *Biometrika,* 51 : 153-167.

50 HUBER, P. J. (1972) Robust statistics : a review. *Annals of Mathematical Statistics,* 43 : 1046.

51 *Loc. cit.*

say, 5 to 10% at the ideal model) to safeguard against ill effects caused by small deviations from it . . . ». He continues, stating that « . . . although I am happy if the procedure performs well under large deviations, I do not really care — inferences based upon a grossly wrong statistical model may have little physical significance. » [52]

The implications of Huber's arguments are clear. If robustness is to have operational utility, those using the concept must make explicit the manner in which it is being used. Even better, researchers should strive towards some universally acceptable goal out of those identified in order to facilitate validation and comparability of research. And, any such definitions must go beyond the shape of distributions. As Huber observes, robustness should also include insensitivity to grouping effects and the like. Some estimates like the sample median, commonly accepted as robust, are not robust relative to grouping effects. Furthermore, he notes that other aspects, such as deviations from independence, exist about which little is known concerning robustness.

The use of robustness by two geographers is now considered in order to reach some conclusions about the way in which the term is used in geographical writing and to determine the utility of the concept in geographical research. When discussing the t distribution and its underlying assumptions, Harvey has commented that

> This test is very powerful (in the statistical sense of power), but the strong assumptions involved can rarely be fulfilled in geography. In some cases we shall be able to show that the necessary conditions are met in the data, but in other cases we are forced to assume their existence. Fortunately, the 't' test has proved to be fairly robust, working even when all the conditions are not strictly met. [53]

King provides the second example. During a discussion of nonparametric statistics, he writes that

> These are distinguished usually by the fact that they do not specify any parameters and, related to this, by the fact that they do not assume any particular form of distribution from which the sample is drawn. In the « classical » approach to statistical inference . . ., much of the theory is derived on the assumption that normal probability distributions are involved. Although certain tests are robust, and not very sensitive to this assumption being violated, it often is preferable to use a distribution-free test in such situations. [54]

In these two sources well known to most geographers, robustness is thus used to suggest that certain parametric techniques may be used when their assumptions are not satisfied. King indicates that distribution-free or non-

---

[52] *Ibid.* 1047.

[53] HARVEY, D. (1969) *Explanation in Geography.* London. Edward Arnold : 280.

[54] KING, L. J. (1969) *Statistical analysis in Geography.* Englewood Cliffs, Prentice-Hall : 83.

parametric tests may be used as well. The concern in the remaining part of this sections is to determine whether robustness is such a useful alternative.

Bradley has presented a strong argument regarding the abuse of robustness.[55] In his view the Normal Mystique is rivaled only by the Myth of Robustness concerning vast overgeneralizations. With both concepts, he believes that « . . . a kernel of truth has been magnified into a mountain of error ».

With regard to robustness, the kernel of truth is that most parametric tests experience impressively little distortion as a result of fairly large violations of assumptions. As examples, he cites the one sample t test for which non-normality has less and less affect as sample size increases. The two sample t test is also noted as being insensitive to heterogeneity as long as other assumptions are satisfied and the two sample sizes are approximately equal. Bradley emphasizes, however, that the conditions which cause a test to be robust are highly idiosyncratic. To illustrate, he points out that the two sample t test is not robust against unequal variances when the sample sizes are not the same. These examples are confirmed by the literature cited previously in this section, especially the work of Boneau.

Given the idiosyncraticity which exists, Bradley contends that the mountain of error « . . . consists in heroic generalizations transcending qualifications and unfettered by definitions. » In this context, his objections against the Myth of Robustness focus on statements such as « the — test is robust against the — assumption » or even worse « the — test is robust ». Since both Harvey and King use robustness in the manner of which Bradley is sharply critical, it is worth following his argument in support of this indictment.

Bradley feels that the above types of statements are unacceptable because they represent semantic chaos. He contends, and is certainly supported by Huber, that no commonly accepted definition exists as to what comprises robustness. Consequently, no criteria have been developed to differentiate between a state of robustness and nonrobustness. With this background, Bradley argues that a reader can only know what the term is supposed to mean by being capable of mind-reading the original investigator. Even worse, according to Bradley, is that the type of statement used by King does not indicate the extent of the violation against which the test is robust. But, in Bradley's words, « . . . the 'amount of robustness' tends to be dependent upon 'the amount of violation'. Thus the only relevant variable mentioned in the statement is not quantified ».

Further problems exist. Bradley argues that robustness against a specific assumption is related to other factors involved in hypothesis testing. These factors do not influence the nature of the Type I error when assump-

[55] BRADLEY. *Op. cit.* 24-43.

tions are met. On the other hand, when assumptions are violated these other factors interact with the assumption violation to jointly influence the Type I error. Several categories of factors exist. One type relates to factors associated exclusively with testing, such as the location of the rejection region (one-tailed or two-tailed test) and the size of the significance level. A second category involves factors concerned with the sampling procedure. Specific items are the minimum sample size, relative and absolute sample size, and the total number of samples upon which the test is based. The third category incorporates factors involving the populations, such as which populations yielded which samples. The insidious aspect, Bradley discovered in his empirical work, is that « Not only do these factors tend to interact with violations of assumptions, but they also display a strong tendency to interact with each other . . . ». The result is that if the 'degree of violation' is held constant but the three types of factors are varied, the robustness of a test will vary.

A confounding aspect of these factors being combined with assumption violations is that the inter-relationships become exceedingly complex. It thereby becomes difficult to anticipate the impact of a factor on *a priori* grounds. Depending upon the particular combination of factors and assumptions, a given violation of an assumption may have a negligible or devastating effect. In fact, the complexity of the interactions may produce completely unexpected results. Thus, one outcome of Bradley's empirical work was that the robustness of a two (equal-sized) sample t-test under a specified violation of normality was greatly increased by introducing a violation to the homogeneity assumption.

Concluding, Bradley emphasizes the following points. First, no objective robustness-nonrobustness dichotomy exists. Instead, the investigator encounters a continuum of degrees of robustness. Second, degree of robustness or nonrobustness is not simply a function of degree of assumption violation. A multiplicity of factors associated with statistical testing procedures interact with assumptions in a complex and unpredictable manner.

The outcome of these findings for the type of statement used by Harvey and King is significant. In order to convert « the — test is robust against the — assumption » into an operational statement about robustness, Bradley indicates that the researcher would have to include three qualifications. First, a quantitative definition of robustness would be needed. Second, a complete statement concerning the extent of assumption violation would be required. And third, a thorough statement of the exact sampling and test conditions should be included. In Bradley's words

> If all this were done, the statement would be so particularistic as to have little general appeal, which perhaps explains why the type of statement quoted survives in its amorphous, undefined, unqualified form. It also

> explains why that form is so completely inaccurate and so utterly meaningless. [56]

The conclusion to be drawn from this discussion is that at the moment robustness does not offer the perfect solution when assumptions are not met during statistical analysis. Huber and Bradley both argue convincingly that the term has yet to be defined operationally. Bradley has also demonstrated that a range of variables must be considered during any effort to refine the concept. Until the concept becomes more sophisticated, it can only be urged that geographers do not abuse the term or rely upon it for a justification which it seems incapable of providing.

CONCLUSIONS AND DISCUSSIONS

This paper has suggested that the geographer contemplating the use of inferential statistical tests faces a number of decisions. First, the investigator must determine the assumptions which underlay the statistical model. Independence, normality, homogeneity and linearity are identified as basic assumptions which are associated with many statistical tests. Second, it is necessary to develop criteria against which decisions may be made as to whether assumptions are satisfied. This stage, it is suggested, requires each individual to develop personally-satisfying criteria which may vary as a result of the nature of the problem, consequences of committing errors, availability of external checks and experience of the investigator. Whatever the criteria used, they should be made explicit and presented for the scrutiny of others. Such a procedure could facilitate validation and comparison of research results.

The third stage requires the researcher to resolve situations in which assumptions are not met. Three alternatives are suggested. Transformations allow non-normal data to be normalized, non-linear data to be made linear, and heterogeneous data to be made homogeneous. Thus, by transforming data the researcher may be able to satisfy assumptions. This gain is often at the expense, however, of making an implicit assumption about the relationships underlying variables and leading to difficulty in interpretting test results. Furthermore, transformations only rarely have a theoretical basis, and must be used with great care if inferences are to be made about processes or patterns associated with spatial activities or man-environment relationships.

Nonparametric or distribution-free tests present a second alternative. While such tests often are unable to cope with higher-order interactions and multivariate designs, they are flexible and offer considerable opportunity in analysis. Many geographers and others have misunderstood the concept of power and power efficiency, and as a result have concluded that nonparametric tests are inferior or wasteful compared to parametric tests. The

---

[56] *Ibid.* 43.

argument here has attempted to clarify an unfortunate misunderstanding about power and the significance of this concept when selecting a statistical technique. It is concluded that the use of power efficiency has led to an unnecessary and unfortunate reluctance to use nonparametric or distribution-free methods.

The third alternative considered involves the concept of robustness. It is demonstrated from the statistics literature that this notion, while conceptually attractive, has not been operationalized completely. Consensus is still lacking about a definition for the term. Moreover, there still is not adequate appreciation concerning the need to specify the degree of assumption violation involved nor the significance of other hypothesis-testing factors in determining the degree of robustness or nonrobustness of a test. With this background, it is suggested that some geographers have, perhaps inadvertently, abused the concept in their research. It is recommended that at this moment the concept is not an adequate justification for proceeding with statistical testing when data do not meet assumptions.

A new point deserves mention in concluding. Variability or dispersion in a set of observations may arise from several different sources. Anscombe has identified three possible sources of variability : inherent variability, measurement error and execution error. [57] Hampel specifies similar sources by noting rounding of observations, occurrence of gross errors, and the model itself only approximating the underlying chance mechanism. [58] The important aspect to stress is that the investigator should only contemplate use of transformations or nonparametric methods if evidences suggests dispersion is a function of inherent variability or related factors. If the dispersion is a function of Anscombe's measurement and execution errors, or Hampel's gross errors, any form of data manipulation will have little theoretical or practical importance.

Bruce MITCHELL
Department of Geography
University of Waterloo
Waterloo, Ontario

---

[57] ANSCOMBE, F. J. (1960) Rejection of outliers. *Technometrics,* 2 : 123-124.

[58] HAMPEL, F. R. (1971) A general qualitative definition of robustness. *Annals of Mathematical Statistics.* 42 : 1887.