Psychological Bulletin 1989, Vol. 105. No.1, 156-166

The Unicorn, The Normal Curve, and Other Improbable Creatures

Theodore Micceri[1]
Department of Educational Leadership
University of South Florida

An investigation of the distributional characteristics of 440 large-sample achievement and psychometric measures found all to be significantly nonnormal at the alpha .01 significance level. Several classes of contamination were found, including tail weights from the uniform to the double exponential, exponential-level asymmetry, severe digit preferences, multimodalities, and modes external to the mean/median interval. Thus, the underlying tenets of normality-assuming statistics appear fallacious for these commonly used types of data. However, findings here also fail to support the types of distributions used in most prior robustness research suggesting the failure of such statistics under nonnormal conditions. A reevaluation of the statistical robustness literature appears appropriate in light of these findings.

1 During recent years a considerable literature devoted to robust statistics has appeared. This research reflects a growing concern among statisticians regarding the robustness, or insensitivity, of parametric statistics to violations of their underlying assumptions. Recent findings suggest that the most commonly used of these statistics exhibit varying degrees of nonrobustness to certain violations of the normality assumption. Although the importance of such findings is underscored by numerous empirical studies documenting nonnormality in a variety of fields, a startling lack of such evidence exists for achievement tests and psychometric measures. A naive assumption of normality appears to characterize research involving these discrete, bounded, measures. In fact, some contend that given the developmental process used to produce such measures, "a bell shaped distribution is guaranteed" (Walberg, Strykowski, Rovai, & Hung, 1984, p. 107). This inquiry sought to end the tedious arguments regarding the prevalence of normal-like distributions by surveying a large number of real-world achievement and psychometric distributions to determine what distributional characteristics actually occur.

2 Widespread belief in normality evolved quite naturally within the dominant reductionist religio-philosophy of the 19th century. Early statistical researchers such as Gauss sought some measure to estimate the center of a sample. Hampel (1973) stated,

> Gauss. . . introduced the normal distribution to suit the arithmetic mean. . . and. . . developed his statistical theories mainly under the criterion of mathematical simplicity and elegance. (p. 94)

---

1. The author holds a joint appointment with the Department of Educational Leadership, College of Education, University of South Florida, and with the Assistant Dean's Office, College of Engineering, Center for Interactive Technologies, Applications, and Research. More complete tables are available from the author for postage and handling costs.
Correspondence concerning this article should be addressed to Theodore Micceri, Department of Educational Leadership, University of South Florida, FAO 296, Tampa, Florida 33620.

3 Certain later scientists, seduced by such elegance, may have spent too much time seeking worldly manifestations of God:

> I know of scarcely anything so apt to impress the imagination as the wonderful form of cosmic order expressed by the "Law of Frequency of Error." The law would have been personified by the Greeks and deified, if they had known of it. It reigns with serenity and in complete self-effacement amidst the wildest confusion. (Galton, 1889, p. 66)

4 Although Galton himself recognized the preceding to hold only for homogeneous populations (Stigler, 1986), such attributions to deity continue to appear in educational and psychological statistics texts:

> It is a fortunate coincidence that the measurements of many variables in all disciplines have distributions that are good approximations of the normal distribution. Stated differently, "God loves the normal curve!" (Hopkins & Glass, 1978, p. 95)

5 Toward the end of the 19th century, biometricians such as Karl Pearson (1895) raised questions about the prevalence of normality among real-world distributions. Distrust of normality increased shortly thereafter when Gosset's (Student, 1908) development of the *t* test, with its strong assumptions, made statisticians of that time "almost over-conscious of universal non-normality" (Geary. 1947, p. 241). During the 1920s, however, an important change of attitude occurred

> following on the brilliant work of R. A. Fisher who showed that, when universal normality could be assumed, inferences of the widest practical usefulness could be drawn from samples of any size. Prejudice in favour of normality returned in full force. . . and the importance of the underlying assumptions was almost forgotten. (Geary, 1947, p. 241)

6 The preceding illustrates both trends in attitudes toward normality and the influence of R. A. Fisher on 20th-century scientists. Today's literature suggests a trend toward distrust of normality; however, this attitude frequently bypasses psychometricians and educators. Interestingly, the characteristics of their measures provide little support for the expectation of normality because they consist of a number of discrete data points and [page 157] because their distributions are almost exclusively multinomial in nature. For multinomial distributions, each possible score (sample point) is itself a variable, and correlations may exist among each variable score/sample point. Thus, an extremely large number of possible cumulative distribution functions (cdfs) exist for such distributions defined by the probability of the occurrence for each score/sample point (Hastings & Peacock, 1975, p. 90). The expectation that a single cdf (i.e., Gaussian) characterizes most score distributions for such measures appears unreasonable for several reasons. Nunnally (1978, p. 160) identifies an obvious one; "Strictly speaking, test scores are seldom normally distributed." The items of a test must correlate positively with one another for the measurement method to make sense. "Average correlations as high as .40 would tend to produce a distribution that was markedly flatter than the normal" (Nunnally, 1978, p. 160). Other factors that might contribute to a non-Gaussian error distribution in the population of interest include but are not limited to (a) the existence of undefined subpopulations within a target population having different abilities or attitudes, (b) ceiling or floor effects, (c) variability in the difficulty of items within a measure, and (d) treatment effects that change not only the location parameter and variability but.also the shape of a distribution.

7 Of course, this issue is unimportant if statistics are truly robust; however, considerable research suggests that parametric statistics frequently exhibit either relative or absolute nonrobustness in the presence of certain nonnormal distributions. The arithmetic mean has not proven relatively robust in a variety of situations; Andrews et al. (1972), Ansell (1973), Gastwirth and Rubin (1975), Wegman and Carroll (1977), Stigler (1977), David and Shu (1978), and Hill and Dixon (1982). The standard deviation, as an estimate of scale, proves relatively inefficient given only 18/100 of 1% contamination (Hampel, 1973). Others who found the standard deviation relatively nonrobust include Tukey and McLaughlin (1963), Wainer and Thissen (1976), and Hettmansperger and McKean (1978). Kowalski (1972) recommends against using

the Pearson product moment coefficient unless (X, Y) is "very nearly normal" because of both nonrobustness and interpretability. Wainer and Thissen (1976) contend that nothing would be lost by immediately switching to a robust alternative, r_t.

8 A large, complex literature on the robustness of parametric inferential procedures suggests that with the exception of the one-mean *t* or *z* tests and the random-effects analysis of variance (ANOVA), parametric statistics exhibit robustness or conservatism with regard to alpha in a variety of nonnormal conditions given large and equal sample sizes. Disagreement exists regarding the meaning of large in this context (Bradley, 1980). Also, several reviews suggest that when *n*s are unequal or samples are small, this robustness disappears in varying situations (Blair, 1981; Ito, 1980; Tan, 1982). In addition, robustness of efficiency (power or beta) studies suggest that competitive tests such as the Wilcoxon rank-sum exhibit considerable power advantages while retaining equivalent robustness of alpha in a variety of situations (Blair, 1981; Tan, 1982).

9 Although far from conclusive, the preceding indicate that normality-assuming statistics may be relatively nonrobust in the presence of non-Gaussian distributions. In addition, any number of works asserting the nonnormality of specific distributions and thereby the possible imprecision of statistical procedures dependent on this assumption may be cited (Allport, 1934; Andrews et al., 1972; Bradley, 1977, 1982; Hampel, 1973; E. S. Pearson & Please, 1975; K. Pearson, 1895; Simon, 1955; Stigler, 1977; Tan, 1982; Tapia & Thompson, 1978; Tukey & McLaughlin, 1963; Wilson & Hilferty, 1929). Despite this, the normality assumption continues to permeate both textbooks and the research literature of the social and behavioral sciences.

10 The implications of the preceding discussion are difficult to assess because little of the noted robustness research deals with real-world data. The complexity and lack of availability of real-world data compels many researchers to simplify questions by retreating into either asymptotic theory or Monte Carlo investigations of interesting mathematical functions. The eminent statistical historian Stephen Stigler (l977), investigating 18th-century empirical distributions, contended, "the present study may be the first evaluation of modern robust estimators to rely on real data" (p. 1070). Those few researchers venturesome enough to deal with real data (Hill & Dixon, 1982; Stigler, 1977; Tapia & Thompson, 1978) report findings that may call much of the above-cited robustness literature into question; (a) Real data evidence different characteristics than do simulated data; (b) statistics exhibit different properties under real-world conditions than they do in simulated environments; and (c) causal elements for parametric nonrobustness tend to differ from those suggested by theoretical and simulated research.

11 In an attempt to provide an empirical base from which robustness studies may be related to the real world and about which statistical development may evolve, the current inquiry surveyed specific empirical distributions generated in applied settings to determine which, if any, distributional characteristics typify such measures. This research was limited to measures generally avoided in the past, that is, those based on human responses to questions either testing knowledge (ability/achievement) or inventorying perceptions and opinions (psychometric).

12 The obvious approach to classifying distributions, à la K. Pearson (1895), Simon (1955), Taillie, Patil, and Baldessari (1981), and Law and Vincent (1983), is to define functionals characterizing actual score distributions. Unfortunately, this approach confronts problems when faced with the intractable data of empiricism. Tapia and Thompson (1978) in their discussion of the Pearson system of curves contend that even after going through the strenuous process of determining which of the six Pearson curves a distribution appears to fit, one cannot be sure either that the chosen curve is correct or that the distribution itself is actually a member of the Pearson family. They suggest that one might just as well estimate the density function itself. Such a task, although feasible, is both complex and uncertain. Problems of identifiability exist for mixed distributions (Blischke, 1978; Quandt & Ramsey, 1978; Taillie et al., 1981), in which the

specification of different parameter values can result in identical mixed distributions, even for mathematically tractable two-parameter distributions such as the Gaussian. Kempthorne (1978) argues that

> "almost all" distributional problems are insoluble with a discrete sample space, notwithstanding the fact that elementary texts are replete with finite space problems that are soluble. (p. 12)

13 [page 158] No attempt is made here to solve the insoluble. Rather, this inquiry attempted, as suggested by Stigler (1977), to determine the degree and frequency with which various forms of contamination (e.g., heavy tails or extreme asymmetry) occur among real data. Even the comparatively simple process of classifying empirical distributions using only symmetry and tail weight has pitfalls. Elashoff and Elashoff (1978), discussing estimates of tail weight, note that "no single parameter can summarize the varied meanings of tail length" (p. 231). The same is true for symmetry or the lack of it (Gastwirth, 1971; Hill & Dixon, 1982). Therefore, multiple measures of both tail weight and asymmetry were used to classify distributions.

14 As robust measures of tail weight, *Q* statistics (ratios of outer means) and *C* statistics (ratios of outer percentile points) receive support. Hill and Dixon (1982), Elashoff and Elashoff(1 978), Wegman and Carroll (1977), and Hogg (1974) discuss the *Q* statistics, and Wilson and Hilferty (1929), Mosteller and Tukey (1978), and Elashoff and Elashoff (1978) discuss the *C* statistics.

15 As a robust measure of asymmetry, Hill and Dixon (1982) recommend Hogg's (1974) Q_2 . However, Q_2 depends on contamination in the tails of distributions and is not sensitive to asymmetry occurring only between the 75th and 95th percentiles. An alternative suggested by Gastwirth (1971) is a standardized value of the population mean/median interval. In the symmetric case, as sample size increases, the statistic should approach zero. In the asymmetric case, as sample size increases, the statistic will tend to converge toward a value indicating the degree of asymmetry in a distribution.

Method

16 Two problems in obtaining a reasonably representative sample of psychometric and achievement/ability measures are (a) lack of availability and (b) small sample sizes. Samples of 400 or greater were sought to provide reasonably stable estimates of distributional characteristics. Distributions, by necessity, were obtained on an availability basis. Requests were made of 15 major test publishers, the University of South Florida's institutional research department, the Florida Department of Education, and several Florida school districts for ability score distributions in excess of 400 cases. In addition, requests were sent to the authors of every article citing the use of an ability or psychometric measure on more than 400 individuals between the years 1982 and 1984 in *Applied Psychology, Journal of Research in Personality, Journal of Personality, Journal of Personality Assessment, Multivariate Behavioral Research, Perceptual and Motor Skills, Applied Psychological Measurement, Journal of Experimental Education, Journal of Educational Psychology, Journal of Educational Research,* and *Personnel Psychology.* A total of over 500 score distributions were obtained, but because many were different applications of the same measure, only 440 were submitted to analysis.

17 Four types of measures were sampled separately: general achievement/ability tests, criterion/mastery tests, psychometric measures, and, where available, gain scores (the difference between a pre- and post-measure).

18 For each distribution, three measures of symmetry/asymmetry were computed: (a) M/M intervals (Hill and Dixon, 1982), defined as the mean/median interval divided by a robust scale estimate(1.4807 multiplied by one-half the interquartile range), (b) skewness, and (c) Hogg's (1974) Q_2 , where

$$Q_2 = [U(05) - M(25)] \ / \ [M(25) - L(05)]$$

where U(alpha)[M(alpha), U(alpha)] is the mean of the upper (middle, lower) [(N + 1)alpha] observations. The inverse of this ratio defines Q_2 for the lower tail.

19 Two different types of tail weight measure were also computed: (a) Hogg's (1974) Q and Q_1, where

$$Q = [U(05) — L(05)] / [U(50)- L(50)]$$
$$Q_1 = [U(20) — L(20)] / [U(50)- L(50)]$$

and (b) C ratios of Elashoff and Elashoff (1978): C_{90}, C_{95}, and $C_{97.5}$ (the ratio of the 90th, 95th, and 97.5th percentile points, respectively, to the 75th percentile point).[1] The Q statistics are sensitive to relative density and the C statistics to distance (between percentiles). Kurtosis, although computed, was not used for classification because of interpretability problems.

20 Criterion values of contamination were determined for these measures using tabled values for symmetric distributions (Elashoff& Elashoff, 1978) and simulated values for asymmetric distributions. Table 1 shows five cut points defining six levels of tail weight (uniform to double exponential) and three cut points defining four levels of symmetry or asymmetry (relatively symmetric to exponential).

Table 1. Criterion Values for Measures of Tail Weight and Symmetry

Distribution	Tail weight					Symmetry/asymmetry		
	C97.5	C95	C90	Q	Q1	Skewness	mn/mdn	Q2
				Expected Values				
Uniform	1.90	1.80	1.60	1.90	1.60	0.00	0.00	1.00
Gaussian	2.90	2.40	1.90	2.58	1.75	0.00	0.00	1.00
Double exponential	4.30	3.30	2.30	3.30	1.93	2.00	0.37	4.70
				Cut Points				
Uniform	1.90	1.80	1.60	1.90	1.60	—	—	—
Below Gaussian	2.75	2.30	1.85	2.50	1.70	—	—	—
Moderate contamination	3.05	2.50	1.93	2.65	1.80	0.31	0.05	1.25
Extreme contamination	3.90	2.80	2.00	2.73	1.85	0.71	0.18	1.75
Double exponential	4.30	3.30	2.30	3.30	1.93	2.00	0.37	4.70

21 Cut points were set arbitrarily, and those defining moderate contamination of either tail weight or asymmetry were selected only to identify distributions as definitely non-Gaussian. The moderate contamination cut points (both symmetric and asymmetric) were set at 5% and 15% contamination on the basis of the support for the alpha trimmed mean and trimmed t in the research literature. Moderate contamination (5%, 2 sd) represents at least twice the expected observations more than 2 standard deviations from the mean, and extreme contamination (15%, 3sd) represents more than 100 times the-expected observations over 3 standard deviations from the mean. Distributions were placed in that category defined by their highest valued measure.

22 Two thousand replications of each classification statistic were computed to investigate sampling error for samples of size 500 and 1,000 for simulated Gaussian, moderate, extreme, and exponential contaminations (Table 1) using International Mathematical and Statistical Library subprograms GGUBS, GGNML, and GGEXN. Only slight differences occurred between sample sizes 500 and 1,000. Each statistic was at expectation for the Gaussian (50% above and 50% below cut). Results for asymmetric conditions indicate

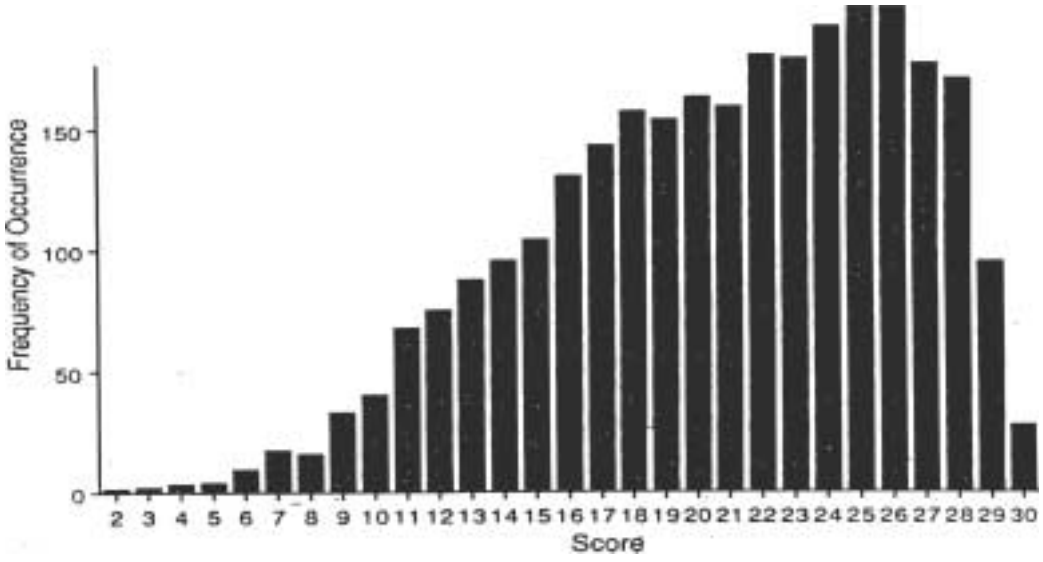
1. Because score distributions did not have a mean of zero, in order to compute percentile ratios it was necessary to subtract the median from each of the relevant percentile points and use the absolute values of the ratios.

that cut points for moderate contamination underestimate nonnormality, with 70.4% (skewness), 81.2% (Q_2), and 72.2% (M/M) of the simulated statistics falling below cut values at sample size 1,000. For extreme asymmetric contamination, simulated values closely fit expectations. However, for the exponential distribution, skewness cut points underestimate contamination (62% below cut), whereas those for Q_2 and M/M overestimate contamination (35% and 43%, respectively, below cut) for sample size 1,000. Among tail weight measures, the most variable estimate ($C_{97.5}$) showed considerable precision for the most extreme distribution (exponential), placing 45% of its simulated values below expected for sample size 1,000. This suggests that one might expect some misclassifications among distributions near the cut points for moderate and exponential asymmetry, with relative precision at other cut values.

23 Figure 1 shows a light-tailed, moderately asymmetric distribution as categorized by the preceding criteria.

24 Multimodality and digit preferences also present identifiability problems for distributions other than the strict Gaussian. Therefore, arbitrary but conservative methods were used to define these forms of contamination. Two techniques, one objective and one subjective, were used to identify modality. First, histograms of all distributions were re- [page 159] viewed, and those clearly exhibiting more than a single mode were classified as such. Second, during computer analysis, all sample points occurring with a frequency at least 80% of that of the true mode (up to a maximum of five) were identified, and the absolute distance between adjacent modes was computed. Distances greater than two thirds (.667) of a distribution's standard deviation ware defined as bimodal. If more than one distance was this great, the distribution was defined as multimodal. In general, the two techniques coincided during application.

Figure 1: A light-tailed, moderately asymmetric distribution (n = 3,152).



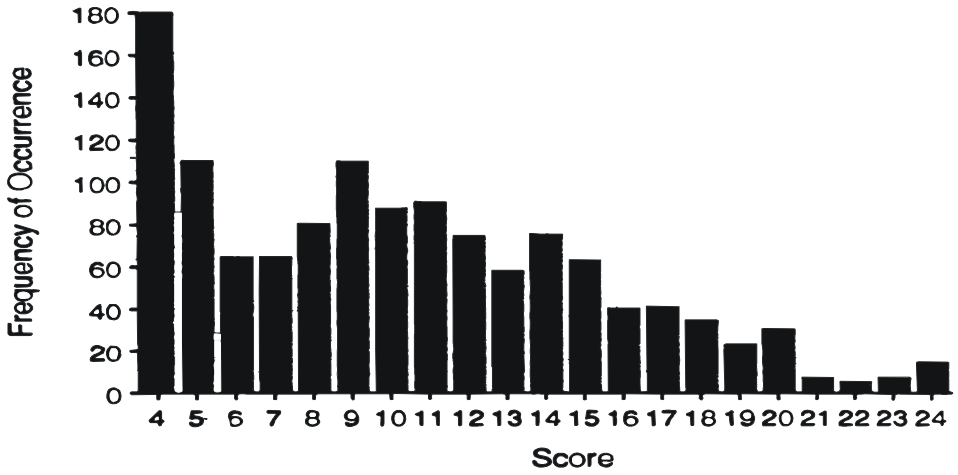
25 Digits were defined as preferred if they occurred at least 20 times and if adjacent digits on both sides had fewer than 70% or greater than 130% as many cases. A digit preference value was computed by multiplying the number of digits showing preference by the inverse of the maximum percentage of preference for each distribution. A digit preference value exceeding 20 (at least four preferred digits with a maximum of 50% preference) was defined as lumpy. In addition, perceived lumpiness was identified. Figure 2 depicts a psychometric distribution that required a perceptual technique for classification as either lumpy or multimodal. This distribution consists of at least two and perhaps three fairly distinct subpopu1ations.

Results

Sample

26 Four hundred and forty distributions were submitted to analysis. Two hundred and sixty-five of these distributions came from journal articles or researches of various types, 30 from national tests, 64 from statewide tests, and 65 from districtwide tests. Seventeen distributions of college entrance and Graduate Record Examination (GRE) scores came from the University of South Florida's admission files.

Figure 2: An asymmetric, lumpy, multimodal distribution (n = 1,258).



27 [page 160] The 231 ability distributions were derived from 20 different test sources (e.g., Comprehensive Test of Basic Skills; CTBS) and 45 different populations. The 125 psychometric distributions included 20 types of measures responded to by 21 different populations. The 35 criterion measures were all part of the Florida State Assessment Program (teacher and student), two test sources responded to by 13 different populations. The 49 gain scores resulted from 5 test sources and 10 different populations.

28 Among ability measures, major sources included the California Achievement Tests, the Comprehensive Assessment Program, the CTBS, the Stanford Reading tests, tests produced by the Educational Testing Service for a beginning teacher study in California, the Scholastic Aptitude Tests, the College Board subject area aptitude tests, the American College Test, the GRE, a series of tests produced by Science Research Associates, several aptitude-tests produced by Project Talent, the Henman Nelson IQ scores from the Wisconsin Longitudinal Study of High School Seniors, the Performance Assessment in Reading of McGraw-Hill, two scores produced by the International Association for the Evaluation of Educational Achievement Student Achievement Study of 1970-1974, and 15 tests representing districtwide, teacher made, textbook-produced, and composite scores created for specific studies.

29 Psychometric measures included: Minnesota Multiphasic Personality Inventory scales; interest inventories; measures of anger, anxiety curiosity, sociability, masculinity/femininity, satisfaction, importance, usefulness, quality, and locus of control; and two measures difficult to categorize, the Mallory test of visual hallucinations and a measure of the degree to which one's partner exerts force to obtain sex.

30 Criterion/mastery test results for students in mathematics and communications skills at the 3rd, 5th, 8th, 10th, and 11th grades were obtained from the Florida State Assessment Program. For adults, Florida Teacher Certification Examination distributions were obtained for reading, writing, mathematics, and-professional education.

31 Sample sizes for the distributions were 190-450 (10.8%), 460-950 (19.8%), 1,000-4,999 (55.1%), and 5,000 to 10,893 (14.3%). Approximately 90% of the distributions included 460 or more cases and almost 70% included 1,000 or more. Subject areas for achievement measures included language arts, quantitative arts/logic, sciences, social studies/history, and skills such as study skills, grammar, and punctuation. Grade/age groupings included 30.5% from grades K-6, 20% from grades 7-9, 18.4% from grades 10-12, 9% from college students, and 22% from adults.

32 Most distributions had sample spaces of between 10 and 99 scale points (83.3%). Fifty-five distributions (12.5%) had sample spaces of fewer than 10 scale points, and 19 distributions (4.3%) had sample spaces greater than 99 scale points.

Measures of Tail Weight and Asymmetry

33 On the basis of the criteria in Table 1, Table 2 shows that 67 (15.2%) of the 440 distributions had both tails with weights at or about the Gaussian, 216 (49.1%) had at least one extremely heavy tail, and 79 (18%) had both tail weights less than the Gaussian. Among ability measures, the percentages were similar with 45 (19.5%) having both tail weights at or about the Gaussian, 133 (57.6%) having at least one heavy tail, and 53 (22.9%) having both tails less than the Gaussian. Among psychometric measures, 17 (13.6%) had tail weights near the Gaussian, 82 (65.6%) had at least one moderately heavy tail, and 26(20.8%) had both tail weights less than the Gaussian. All criterion/mastery and 45 (89.8%) of the gain score distributions exhibited at least one tail weight greater than that expected at the Gaussian. Five gain scores (l0.2%)had tail weights near the Gaussian.

34 Table 3 shows that among all measures, 125 of the distributions were classified as being relatively symmetric (28.4%), and 135 (30.7%) were classified as being extremely asymmetric. Forty-seven percent of the gain score, 65.8% of the ability/achievement measures, 84.0% of psychometric measures, and 100% of criterion/mastery measures were at least moderately asymmetric. Criterion/mastery and psychometric measures frequently exhibited extreme to exponential asymmetry, 94.3% and 52.0%, respectively. General ability measures tended to be less extreme (15.6% extremely or exponentially asymmetric).

35 Crossing the values for tail weight and symmetry, Table 4 shows that 30 (6.8%) of the 440 distributions exhibit both tail weight and symmetry approximating that expected at the Gaussian and that 21 (48%) exhibited relative symmetry and tail weights lighter than that expected at the Gaussian.

Table 2. Categories of Tail Weight Across Types of Measures, %

Level of symmetric contamination	Achievement (n = 231)	Psychometric (n = 125)	Criterion mastery (n = 35)	Gain score (n = 49)	All types (n = 440)
Uniform	00.0	11.2	00.0	00.0	3.2
Less than Gaussian	22.9	9.6	00.0	00.0	14.8
About Gaussian	19.5	13.6	00.0	10.2	15.2
Moderate	19.9	14.4	8.6	22.5	17.7
Extreme	29.4	28.0	31.4	59.2	32.5
Double exponential	8.2	23.2	60.0	8.2	16.6
Total	100.0	100.0	100.0	100.0	100.0

36 [page 161] Table 5 shows that results were similar for ability measures, with 23 (10.0%) at or about the Gaussian and 20 (8.7%) exhibiting relative symmetry and tail weights less than that expected at the Gaussian.

Table 3. Categories of Asymmetry Across Types of Measures, %

Level of asymmetric contamination	Achievement (n = 231)	Psychometric (n = 125)	Criterion mastery (n = 35)	Gain score (n = 49)	All types (n = 440)
Relatively symmetric	34.2	16.2	00.0	53.1	28.4
Moderate asymmetry	50.2	32.0	5.7	42.9	40.7
Extreme asymmetry	12.6	33.6	37.1	4.1	19.5
Exponential asymmetry	3.0	18.4	57.1	0.0	11.4
Total	100.0	100.0	100.0	100.0	100.0

37 Table 6 shows that 4 psychometric distributions (3.2%) exhibited both relative symmetry and tail weights near the Gauss ian and 39 distributions (3 1.2%) exhibited extreme- to exponential-level tail weight combined with extreme- to exponential-level asymmetry.

38 Table 7 shows that criterion/mastery measures tended to exhibit at least moderate asymmetry (100%) and at least one tail weight at either the extreme or exponential level (91.4%). Twenty (57.2%) of these distributions exhibited asymmetry at or above the exponential.

39 Table 8 shows that gain scores were relatively symmetric to moderately asymmetric with moderate to heavy tail weights (81.6%). Four cases (8.2%) exhibited tail weight at or above the double exponential, and five (10.2%) were at or about the Gaussian. Two distributions (4.1%) exhibited asymmetry greater than the moderate level.

40 Although not used as a classification measure, kurtosis estimates were computed and ranged from -1.70 to 37.37.[1] Ninety-seven percent (35136) of those distributions exhibiting kurtosis beyond the double exponential (3.00) also showed extreme or exponential asymmetry and were frequently characterized by sample spaces of greater than 25 scale points. Almost all distributions having low (negative) kurtoses were at most moderately asymmetric and frequently had small sample spaces. The fourth-moment kurtosis estimate for these distributions correlated r = .78 with the third-moment skewness estimate.

Modality and Digit Preferences

41 Three hundred and twelve (70.9%) distributions were classified as unimodal, 89 (20.2%) as bimodal, and 39 (8.9%) as multimodal. Two hundred and eighteen distributions (49.5%) were defined as relatively smooth and 222 (50.5%) as lumpy. The smoothest distributions were criterion/mastery measures (89%) and gain scores (73%). Psychometric measures tended to be lumpy (6 1.6%), as did general ability measures (54.3%).

1. These are adjusted values at which the expected value at the Gaussian is 0.00 rather than 3.00.

42 The Kolmogorov-Smirnov test of normality (SAS Institute, 1985) found 100% of the distributions to be significantly nonnormal at the .01 alpha level. However, 16 ability measures (6.9%) and 3 gain scores (6.1%) were found to be relatively symmetric, smooth, and unimodal and to have tail weights near those expected at the Gaussian. These 19 distributions (4.3%) may be considered quite reasonable approximations to the Gaussian. No psychometric measures and no criterion/mastery measures were included among these 19 distributions. Sample spaces ranged from 7 to 135 and sample sizes from 346 to 8,092.

Discussion

43 Although not drawn randomly, the 440 distributions coming from some 46 different test sources and 89 different populations should include most types of distributions occurring in applied settings for these measures. Since 60% of all distributions result directly from research and another 33% from state, district, or university scoring programs, they should also represent distributions directly relevant to research, theory development, and decision making.

44 Walberg et al. (1984), on the basis of an impressive literature review, conclude that asymmetry and extremes lying several standard deviations above the main distribution body occur commonly where measures "are less restrictive in range than the typical achievement and attitude scale" (p. 107). The current inquiry shows that even among the bounded measures of psychometry and achievement, extremes of asymmetry and lumpiness are more the rule than the exception. No distributions among those investigated passed all tests of normality, and very few seem to be even reasonably close approximations to the Gaussian. It therefore appears meaningless to test either ability or psychometric distributions for normality, because only weak tests or chance occurrences should return a conclusion of normality. Instead, one should probably heed Geary's (1947) caveat and pretend that "normality is a myth; there never was, and never will be, a normal distribution" (p. 241).

45 The implications of this for many commonly applied statistics are unclear because few robustness studies, either empirical or theoretical, have dealt with lumpiness or multimodality. These findings suggest the need for careful data scrutiny prior to analysis, for purposes of both selecting statistics and interpreting results. Adequate research is available to suggest that most parametric statistics should be fairly robust to both alpha and beta given light tail weights and moderate contaminations. For extreme to exponential asymmetry (52.0% of psychometric measures), one might expect at least the independent means t (given approximately equal ns) and F to exhibit robustness to alpha, if not beta. However, under such conditions, differences between medians may well be a more interesting research question than mean shift for studies seeking information about the middle rather than the tails of a distribution (Wilcox & Charlin, 1986).

46 Normalizing transformations are frequently applied to sus- [page 162]pected departures from symmetry. These, however, should be used with caution, because of problems such as selection and interpretability. For instance, as E. S. Pearson and Please (1975) note regarding log transformations, "There are also pitfalls in interpreting the analysis, if only because the antilog of the mean value of log x is not the mean of x" (p. 239). On this topic, see also Taylor (1985), Games (1984), Hill and Dixon (1982), Bickel and Doksum (1981), Carroll (1979), and Mosteller and Tukey (1978).

Table 4. Tail Weight and Asymmetry for All Distributions

Values of tail weight	Values of asymmetry				Total	
	Near symmetry *n*	Moderate *n*	Extreme *n*	Exponential *n*	*N*	Percentage
Uniform	0	4	5	5	14	3.2
Less than Gaussian	21	33	8	3	65	14.8
Near Gaussian	30	29	7	1	67	15.2
Moderate contamination	30	35	11	2	78	17.8
Extreme contamination	41	64	35	3	143	32.5
Double exponential	3	14	20	36	73	16.6
Total	125	179	86	50	440	—
Percentage	28.4	40.7	19.6	11.4	—	100.0

47 An attempt was made to characterize easily discernable [*sic*] groups of distributions. Patterns occurred consistently for two measures: (a) Gain scores tended to be fairly symmetric (either symmetric or moderately asymmetric) and to have moderate to heavy tails (85.7% of gain score distributions); (b) criterion/ mastery tests tended to be extremely asymmetric (94.3%), with at least one heavy tail (9 1.4%). Fully 85.7% of the criterion/ mastery distributions have at least one heavy tail combined with extreme asymmetry.

48 It proved impossible, however, to typify either general ability/achievement or psychometric measures, both of which tended to distribute throughout the symmetry/tail weight matrix (Tables 5 and 6), while exhibiting varying modalities and digit preferences. Psychometric measures exhibited greater asymmetry (84% were at least moderately asymmetric) and heavier tails (65.6% had at least one moderately heavy tail) than did ability measures.

49 Table 5 suggests that general ability measures tend to exhibit less extreme contamination than do the other measures. None had tail weights at or near the uniform, and only 3.0% exhibited asymmetry at or above that expected for the exponential. However, even if one treats all moderately contaminated cells of Table 5 as reasonable approximations to normality, only 132 general ability distributions (57.1%) would qualify for the title.[1]

50 Table 4 shows that most cells of the tail weight/asymmetry matrix are filled and that counts in each cell tend to remain fairly constant as one moves from light tails to heavy tails or from relative symmetry to extreme asymmetry. Table 4 also shows the poor match between real data and the smooth mathematical functions generally applied in Monte Carlo robustness studies. Distributions exhibiting either extremely heavy tail weights (exponential) or extremely light tail weights (uniform) tend also to be asymmetric. This suggests that simulated studies based on such symmetric mathematical functions as the uniform, logistic, double exponential, Cauchy, and *t* with few degrees of freedom may not represent real-world data to any reasonable extent.

1. Recall that moderate contamination represents at least twice the expected cases more than 2 standard deviations from the mean and not more than 100 times the expected cases more than 3 standard deviations from the mean.

51 The distributions studied here exhibited almost every conceivable type of contamination, including (a) broad classes of tail weight (uniform to double exponential), (b) broad classes of symmetry (quite symmetric to asymmetry greater than that of the exponential), (c) varying modalities (unimodal, bimodal, multimodal), (d) varying types of lumpiness/digit preference, and (e) modes external to the mean/median interval. Also, all ratios of a robust scale estimate to the standard deviation were greater than the 1.00 expected at the normal. This indicates that all distributions exhibit at least some asymmetry (Messick, 1982; K. Pearson, 1895).

52 The great variety of shapes and forms suggests that respondent samples themselves consist of a variety of extremely heterogeneous subgroups, varying within populations on different yet similar traits that influence scores for specific measures. When this is considered in addition to the expected dependency inherent in such measures, it is somewhat unnerving to even dare think that the distributions studied here may not represent most of the distribution types to be found among the true populations of ability and psychometric measures.

53 One might expect treatment effects to create lumpiness, subgroupings, or bi/multimodalities such as those encountered in these data. Although a likely effect, it does not influence these results because the large sample requirement essentially eliminated postmeasures from experimental studies. In those situations in which both pre- and postmeasures were available, almost every case exhibiting lumpiness or bi/multimodality in the postmeasure showed similar characteristics in the premeasure. Figure 3 depicts an interesting and fairly common example of this with an intervening treatment. This premeasure, classified as neither bimodal nor lumpy, appears to include two [page 163] subgroups. One is familiar with the material, approaches the test ceiling and ranges about 12-17. The second is unfamiliar with the material and distributes around 4-7. The unimodal nature of the postmeasure suggests that treatment (a 6-week general biology course) largely eliminated the latter group.

54 To assure that distributions were as homogeneous as possible, all distributions having identified subpopulations that were expected to differ on the measure (e.g., White/non-White, male/female) were separated, and generally only one was submitted to analysis. That distributions still exhibited substantial lumpiness and varying modalities calls to mind the argument Cournot proposed in 1843 that probability is irrelevant to statistics in the social sciences because

> an unlimited number of ways of classifying social data existed and any probability analysis that did not allow for the selection of categories after the collection of data was, in a practical sense, meaning1ess. (Stigler, 1986, P. 197)

Table 5. Tail Weight and Asymmetry for Ability Distributions

Values of tail weight	Values of asymmetry				Total	
	Near symmetry *n*	Moderate *n*	Extreme *n*	Exponential *n*	*N*	Percentage
Uniform	0	0	0	0	0	0.0
Less than Gaussian	20	29	4	0	53	22.9
Near Gaussian	23	19	3	0	45	19.5
Moderate contamination	15	26	5	0	46	19.9
Extreme contamination	18	35	14	1	68	29.4
Double exponential	3	7	3	6	19	8.2
Total	79	116	29	7	231	—
Percentage	34.2	50.2	12.6	30	—	100.0

55 The use of multiple classification measures produced some interesting findings. As with simulated exponential distributions, Q_2 uniquely defined more real-world distributions as beyond the exponential (eight) than did either skewness (six) or M/M (four). Q statistics for tail weight (Q, Q_1,) rarely reached the highest classification value largely because of the prevalence of asymmetry. For C statistics, negative tails were more frequently defined as non-Gaussian than were positive ones. Also, contamination occurred more frequently in the closer tails (C_{10} /C_{90}) than in the farther tails (C_{025} /C_{975}). This suggests that contamination in the tails for these distributions is not evenly distributed, as one might expect for bounded, lumpy populations including undefined subgroups.

56 Some may contend that the use of finite samples does not disprove normality, because as sample size increases, score distributions are attracted to the normal. This type of confusion stems from the fallacious overgeneralization of central limit theorem properties from sample means to individual scores. The central limit theorem states that the sums (or means) of sufficiently large samples from a population satisfying the Lindberg conditions will have an approximately normal distribution. It does not state, however, that the population of scores from which these sample means are drawn is normally distributed (Tapia & Thompson, 1978).

57 As was noted earlier, the implications these findings have for normality-assuming statistics are unclear. Prior robustness studies have generally limited themselves either to computational evaluation of asymptotic theory or to Monte Carlo investigations of interesting mathematical functions. This research [page 164]has been conducted almost exclusively using smooth mathematical functions that have rather extreme tail weights or asymmetry. Such characteristics proved rare among these real-world distributions. Because 50% of these distributions exhibited lumpiness and about two thirds of ability and over four fifths of psychometric measures exhibited at least moderate asymmetry, these appear to be important areas for future study.

Table 6. Tail Weight and Asymmetry for Psychometric Distributions

Values of tail weight	Values of asymmetry					Total	
	Near symmetry *n*	Moderate *n*	Extreme *n*	Exponential *n*		*N*	Percentage
Uniform	0	4	5	5	0	14	11.2
Less than Gaussian	1	4	4	3		12	9.6
Near Gaussian	4	9	3	1		17	13.6
Moderate contamination	6	7	3	2		18	14.4
Extreme contamination	9	12	12	2		35	28.0
Double exponential	0	4	15	10		29	23.2
Total	20	40	42	23		125	—
Percentage	16.0	32.0	33.6	18.4		—	100.0

58 Interestingly, in Andrews et al. (1972, p. 109) there is a small section entitled “Asymmetric Situations” beginning with the caution, “Except in a few instances there may be no reason to believe the underlying distribution is symmetric.” Andrews et al. (1972) investigated the performance of 65 location estimators in the presence of simulated normal populations having 10% asymmetric contamination 2 and 4 standard deviations from the population mean. For both situations at all sample sizes, the arithmetic mean proved

the least variable (best) estimator. These authors, who concluded that the arithmetic mean was the "best" choice as the "worst" estimator among those investigated, fail to mention this finding again because "about unsymmetric situations,. . . we were not able to agree, either between or within individuals, as to the criteria to be used" (Andrews et al., 1972, p. 226). Thus, the arithmetic mean proved most robust (least variable) under asymmetry, the condition found to occur for most (71.6%) distributions investigated here.

Table 7. Tail Weight and Asymmetry for Criterion/Mastery Measures

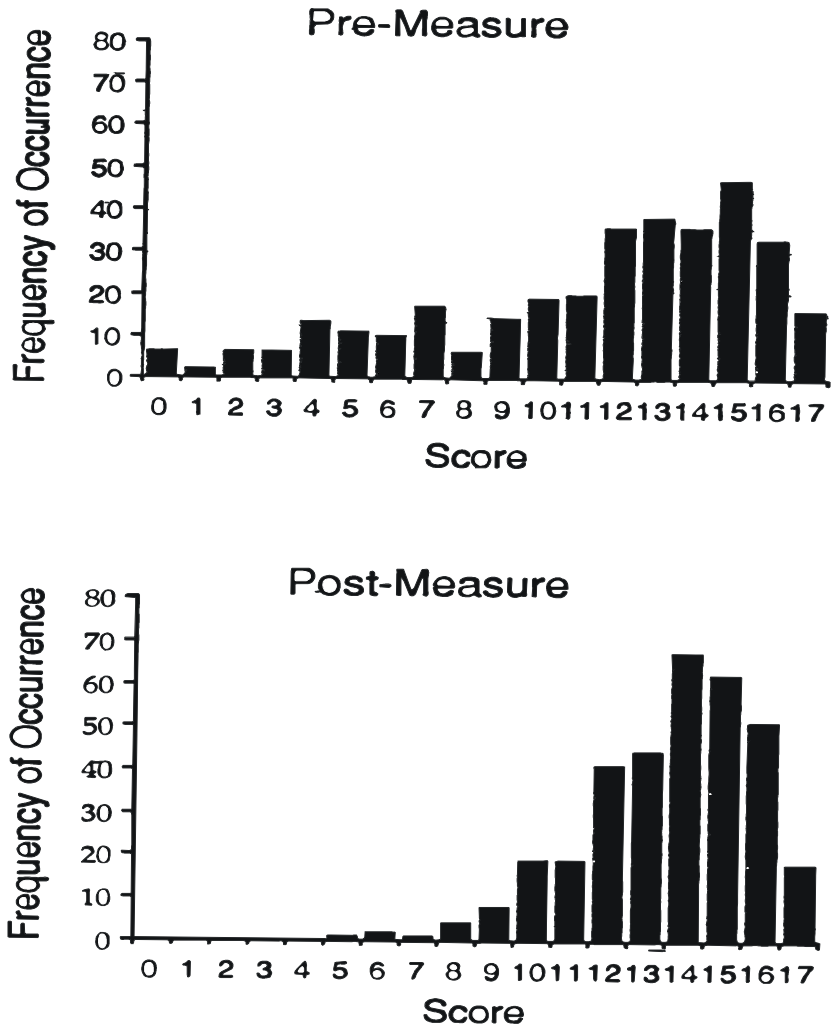
Values of tail weight	Values of asymmetry				Total	
	Near symmetry *n*	Moderate *n*	Extreme *n*	Exponential *n*	*N*	Percentage
Uniform	0	0	0	0	0	0.0
Less than Gaussian	0	0	0	0	0	0.0
Near Gaussian	0	0	0	0	0	0.0
Moderate contamination	0	0	3	0	3	8.6
Extreme contamination	0	2	9	0	11	31.4
Double exponential	0	0	1	20	21	60.0
Total	0	2	13	20	35	—
Percentage	0.0	5.7	37.1	57.2	—	100.0

Table 8. Tail Weight and Asymmetry for Gain Scores

Values of tail weight	Values of asymmetry				Total	
	Near symmetry *n*	Moderate *n*	Extreme *n*	Exponential *n*	*N*	Percentage
Uniform	0	0	0	0	0	0.0
Less than Gaussian	0	0	0	0	0	0.0
Near Gaussian	3	1	1	0	5	10.2
Moderate contamination	9	2	0	0	11	22.5
Extreme contamination	14	15	0	0	29	59.2
Double exponential	0	3	1	0	4	8.2
Total	26	21	2	0	49	—
Percentage	53.1	41.9	4.1	0.0	—	100.0

59 Factors such as these suggest the need (a) to investigate the previous robustness research and determine its appropriateness given the types of contamination found to exist in the real world and (b) to suggest important areas for the investigation of the robustness of various statistics.

Figure 3: Pre- and postmeasures in 10th grade general biology (n = 337).



60 As an example of the first suggestion, the oft-cited works of Boneau (1960, 1962) and two prior studies dealing with small sample space situations are superficially considered. Boneau (1960, 1962) compared the robustness of the Mann-Whitney/ Wilcoxon rank-sum test to that of the t test for samples of size (5, 5), (15, 15), and (5, 15) in the presence of two smooth symmetric distributions (uniform and normal) and one smooth asymmetric distribution (exponential). Among distributions studied here, notwithstanding the fact that all of his distributions were continuous and smooth, although half of these real-world data sets were lumpy and all were discrete, only 38 (8.6%) exhibited both exponential-level tail weight and asymmetry (largely criterion/mastery measures, n = 20), none exhibited symmetric, uniform (rectangular) tail weights, and only 19 (4.3%) can be considered even reasonable approximations to the Gaussian (normal). This does not invalidate his findings but does suggest that almost none of these comparisons occurs in real life. The most obvious differences between Boneau's data and that of the real world are lumpiness and discreteness. Two [page 165] prior studies deal with distributions exhibiting such characteristics in the limited arena of small sample spaces. Hsu and Feldt (1969) found the F to exhibit robustness to alpha for populations with from 3 to 6 scale points (sample space). However, the maximum third-moment skewness included among their populations was .39, and in the current study, among the 43 distributions having sample spaces between 3 and 6, 72.7% exhibited either positive or negative skew greater

than .39. Thus, at least one important distributional characteristic suggests that the findings of Hsu and Feldt may not generalize to the real world of small sample spaces.

61 In a recent study by Gregoire and Driver (1987), the authors investigated several statistics in the presence of 12 varied populations having sample spaces of four or five. Among the 18 distributions in the current study having sample spaces of five or less, 7 (38.8%) exhibited skewness at or greater than .94. Only one population studied by Gregoire and Driver (1987) exhibited asymmetry at or about that level (0.99), and it proved to be one of the worst populations in their article. Specifically, for population IIC, the two-sample parametric confidence interval tended to be conservative to alpha (supporting almost all prior research using equal *n*s). The population mean was outside the .05 confidence interval about the sample mean 75% of the time for samples of size 25. The *F* test for homogeneity of variance was operating at an obtained alpha of about .21 when nominal alpha was .05. And finally, the KS two-sample test was extremely conservative, having an obtained alpha of about .01 when nominal alpha was .05. Unfortunately, this population was not included in their discussion of power. However, from their Table 6, it is interesting to note that the only comparison between two-sample tests in which a substantial power advantage accrues to any test is that between populations IIIA and IA (uniform). In that situation, the van der Waerden test exhibited a considerable power advantage at sample size 10 over both the parametric confidence interval and the Mann-Whitney/Wilcoxon tests. The current study suggests that this specific situation may never arise in practice, because none of the 440 distributions investigated here exhibited both relative symmetry and uniform level tail weights. However, 53 (22.9%) of ability/achievement distributions and 26 (20.3%) of psychometric distributions did have both tails lighter than the Gaussian.

62 Overall, one must conclude that the robustness literature is at best indicative, for at least two reasons: (a) Few prior studies deal with commonly occurring characteristics such as lumpiness and multimodalities, and (b) in some circles (e.g., Andrews et al., 1972), bias against the finding of robustness for parametric statistics may exist.

63 One disturbing finding of this research was a general lack of data availability. Only about 25% of the authors to whom requests were sent reported the ability to produce simple frequency distributions for data reported in their studies. Many different reasons for this inability were noted; however, no matter what the reasons, the situation is somewhat disquieting.

References

Allport, E M. (1934). The J-curve hypothesis of conforming behavior. *Journal of Social Psychology, 5*, 14 1-183.

Andrews, D. E, Bickel, P. J., Hampel, F. R., Huber, P. J., Rogers, W. H., & Tukey, J. W. (1972). *Robust estimates of location survey and advances*. Princeton, NJ: Princeton University Press.

Ansell, M. J. G. (1973). Robustness of location estimators to asymmetry. Applied Statistics, 22, 249-254.

Bickel, P. J., & Doksum, K. A. (1981). An analysis of transformations revisited. *Journal of the American Statistical Association, 76,* 296-311.

Blair, R. C. (1981). A reaction to "Consequences of failure to meet assumptions underlying the fixed effects analysis of variance and covariance." *Review of Educational Research, 51,* 499-507.

Blischke, W. R. (1978). Mixtures of distributions. In W. H. Kruskal and J. M. Tanur (Eds.), *International encyclopedia of statistics* (pp. 174-180). New York: Free Press.

Boneau, C. A. (1960). The effects of violations of assumptions underlying the *t* test. *Psychological Bulletin, 57*, 49-64.

Boneau, C. A. (1962). A comparison of the power of the *U* and *t* tests. Psychological Review, 69, 246-256.

Bradley, J. W. (1977). A common situation conducive to bizarre distribution shapes. *The American Statistician, 31,* 147-150.

Bradley, J. W. (1980). Nonrobustness in *z*, *t*, and *F* tests at large sample sizes. *Bulletin of the Psychonomic Society, 16,* 333-336.

Bradley, J. W. (1982). The-insidious L-shaped distribution. *Bulletin of the Psychonomic Society, 20,* 85-88.

Carroll, R. J. (1979). On estimating variances of robust estimators when the errors are asymmetric. *Journal of the American Statistical Association, 74,* 674-679.

David, H. A., & Shu, V. S. (1978). Robustness of location estimators in the presence of an outlier. In H. A. David (Ed.), *Contributions to survey sampling and applied statistics* (pp. 235-250). New York: Academic Press.

Elashoff, J. D., & Elashoff, R. M. (1978). Effects of errors in statistical assumptions. In W. H. Kruskal and J. M. Tanur (Eds.), *International encyclopedia of statistics* (pp. 229-250). New York: Free Press.

Galton, F. (1889). Natural inheritence. London: Macmillan. Games, P. A. (1984). Data transformations, power, and skew: A rebuttal to Levine and Dunlap. *Psychological Bulletin, 95,* 345-347. [*sic*]

Gastwirth, J. L. (1971). On the sign test for symmetry. *Journal of the American Statistical Association, 166,* 821-823.

Gastwirth, J. L., & Rubin, H. (1975). The behavior of robust estimators on dependent data. *The Annals of Statistics, 3,* 1070-1100.

Geary, R. C. (1947). Testing for normality. *Biometrika, 34,* 209-242.

Gregoire, T. G., & Driver, B. L. (1987). Analysis of ordinal data to detect population differences. *Psychological Bulletin, 101,* 159-165.

Hampel, F. R. (1973). Robust estimation: A condensed partial survey. *Zeitschrzft fur Wahrscheinlichkeitstheorie und Verwandte Gebiete, 27,* 87-104.

Hastings, N. A. J., & Peacock, J. B. (1975). *Statistical distributions: A handbook for students and practitioners.* New York: Wiley.

Hettmansperger, T P., & McKean, J. W. (1978). Statistical inference based on ranks. *Psychometrika, 43,* 69-79.

Hill, M., & Dixon, W. J. (1982). Robustness in real life: A study of clinical laboratory data. *Biometrics, 38,* 377-396.

Hogg, R. V. (1974). Adaptive robust procedures: A partial review and some suggestions for future applications and theory. *American Statistical Association Journal, 69,* 909-927.

Hopkins, K. D., & Glass, G. V. (1978). *Basic statistics for the behavioral sciences.* Englewood Cliffs, NJ: Prentice-Hall.

Hsu, T., & Feldt, L. S. (1969). The effect of limitations on the number of criterion score values on the significance level of the *F* test. *American Educational Research Journal, 6,* 5 15-527.

Ito, P. K. (1980). Robustness of ANOVA and MANOVA test procedures. In P. R. Krishnaiah (Ed.), *Handbook of statistics* (Vol. 6, pp. 199-236). Amsterdam: North-Holland.

Kempthorne, O. (1978). Some aspects of statistics, sampling and randomization. In H. A. David (Ed.), *Contributions to survey sampling and applied statistics* (pp. 11-28). New York: Academic Press.

Kowaiski, C. L (1972). On the effects of non-normality on the distribution of the sample product-moment correlation coefficient. *Applied Statistics, 21,* 1-12.

Law, A. M., Vincent, S. O. (1983). *UNIFIT: An interactive computer package for fitting probability distributions to observed data.* Tucson, AZ: Simulation Modeling and Analysis Company.

Messick, D. M. (1982). Some cheap tricks for making inferences about distribution shapes from variances. *Educational and Psychological Measurement, 42,* 749-758.

Mosteller, F., & Tukey, J. W. (1978). *Data analysis and regression: A second course in statistics.* Boston: Addison-Wesley.

Nunnally, J. C. (1978). *Psychometric theory.* New York: McGraw-Hill.

Pearson, E. S., & Please, N. W. (1975). Relation between the shape of population distribution and the robustness of four simple test statistics. *Biometrika, 62,* 223-241.

Pearson, K. (1895). Contributions to the mathematical theory of evolution: II. Skew variation in homogeneous material. *Philosophical Transactions of the Royal Society Ser. A,* 186, 343-414.

Quandt. R. E., & Ramsey, J. B. (1978). Estimating mixtures of normal distributions and switching regressions. *American Statistical Association Journal, 73*, 730-738.

SAS Institute. (1985). *SAS user's guide: Basics*. Cary, NC: Author.

Simon, H. A. (1955). On a class of skew distribution functions. *Biometrika, 42*, 425-440.

Stigler, S. M. (1977). Do robust estimators work with real data? *The Annals of Statistics, 5*, 1055-1098.

Stigler, S. M. (1986). *The history of statistics: The measurement of uncertainty before 1900*. Cambridge, MA: Belknap Press.

Student. (1908). The probable error of a mean. *Biometrika, 6*, 1-25.

Taillie, C., Patil, G. P., & Baldessari, B. A. (1981). *Statistical distributions in scientific work: Vol. 5. Inferential problems and properties*. Boston: D. Reidel.

Tan, W. Y. (1982). Sampling distributions and robustness oft, F and variance-ratio in two samples and ANOVA models with respect to departure from normality. *Communications in Statistics. A11*, 2485-2511.

Tapia, R. A., & Thompson, J. R. (1978). *Nonparametric probability density estimation*. Baltimore, MD: Johns Hopkins University Press.

Taylor, J. M. G. (1985). Measures of location of skew distributions obtained through Box-Cox transformations. *Journal of the American Statistical Association, 80,* 427-432.

Tukey, J. W., & McLaughlin, D. H. (1963). Less vulnerable confidence and significance procedures for location based on a single sample: Trimming/Winsorization. *Indian Journal of Statistics, 25*, 331-351.

Wainer, H., & Thissen, D. (1976). Three steps toward robust regression. *Psychometrika, 41,* 9-34.

Walberg, H. J., Strykowski, B. E, Rovai, E., & Hung, S. S. (1984). Exceptional performance. *Review of Educational Research, 54*,87-112.

Wegman, E. J., & Carroll, R. J. (1977). A Monte Carlo study of robust estimators of location. *Communications in Statistics, A6*, 795-8l2.

Wilcox, R. R., & Charlin, V. L. (1986). Comparing medians: A Monte Carlo study. *Journal of Educational Statistics, 11*, 263-274.

Wilson, E. B., & Hilferty, M. M. (l929). Note on C. S. Peirce's experimental discussion of the law of errors. *Proceedings of the National Academy of Science, 15,* 120-125.

Received September 14, 1987

Revision received November 30, 1987

Accepted March 22, 1988