Bulletin of the Psychonomic Society
1980, Vol. 16 (5), 333-336

Nonrobustness in Z, t, and F tests at large sample sizes

JAMES V. BRADLEY
New Mexico State University, Las Cruces, New Mexico 88003

The alleged robustness of Z, t, and F tests against nonnormality and, when sample sizes are equal, of t and F tests against heterogeneity as well was investigated in a large-scale sampling study under conditions realistic to experimentation and testing in the behavioral sciences. Factors varied were: population shape (L or bell), σ_1/σ_2 (1/2, 1, or 2), size N of smallest sample (2, 4, 8, 16, 32, 64, 128, 256, 512, or 1,024), N_1/N_2 (1/3, 1/2, 1, 2, or 3), α (.05, .01, or .001), and test tailedness (left, right, or two). In about 25% of the situations investigated, the test failed to meet a very lax criterion for robustness at every examined N value less than 100, and in 8% at every value less than 1,000; no test met the criterion in all of the situations studied before N=512. Robustness was strongly influenced by all of the factors investigated, and interactions among the influencing factors were often strong and complex.

It is often alleged, without further qualification, that the classical Z, t, and F tests on means are robust against nonnormality and that, when sample sizes are equal, t and F are robust against heterogeneity of variance, as well. Although there are highly qualified situations in which the allegation would be supported, as a nearly unqualified generalization it is false.

The origin of the fallacy appears to lie in a series of logical errors including extrapolation downward from infinity to small sample sizes (Ns), ignoring or downplaying contrary evidence, overgeneralizing selected favorable evidence obtained under highly limited conditions, and pure wishful thinking (Bradley, 1978). It can be proved mathematically (Bradley, 1968b, 1976; Scheffé, 1959) that when sample sizes are infinite, the Z, t, and F tests on means are perfectly robust against nonnormality, and, if the samples are of equal as well as infinite size, the two-independent-sample t and F tests are perfectly robust (and the multisample F test is sometimes fairly robust) against heterogeneity of variance, either alone or in combination with nonnormality. This implies that there must be some large finite N at which the robustness that was perfect at infinite N simply becomes nearly perfect; this has encouraged certain mathematicians (e.g., Scheffé, 1959) to claim robustness at "large" N. However, it was not made sufficiently clear that "large" simply means "however large is necessary to produce the desired effect." Unfortunately, "large" seems often to have been interpreted as anything that is not "small," in which case the problem becomes one of identifying the minimum N value above which universal robustness reigns. The fallacy of this approach should have been revealed by experimentation, and empirical sampling studies have often shown very mixed results. However, the robust portion of the results has often been emphasized and overgeneralized to uninvestigated conditions while the unfavorable results (such as rapidly diminishing robustness with diminishing α values) have been downplayed or ignored, sometimes being implicitly attributed to sampling error, and sometimes the partial robustness obtained at the N values actually investigated has emboldened the author to specify a slightly larger sample size at which we shall allegedly be safe (see Boneau, 1960, for examples of all of these practices). Such claims have been echoed and still further liberalized and distorted by writers of textbooks (see Bradley, 1978), often to the point at which the robustness effects that are true at infinite N are now being claimed without reference to absolute sample size. For example, Glass and Stanley (1970, p. 297) state that "Violation of the assumption of normality in the t-test of $H_0: \mu_1 - \mu_2 = 0$ has been shown to have only trivial effects on the level of significance and power of the test and hence should be no cause for concern (Boneau, 1960; Scheffé, 1959, chap. 10) . . . . If n_1 and n_2 are equal, violation of the homogeneous variances assumption is unimportant and need not concern us (Box, 1954a, b; Scheffé, 1959, chap. 10)."

The present study will show that in a variety of realistic cases, sample sizes in excess of 100 or even 1,000 are inadequate to produce the robustness claimed, even though a very liberal criterion for robustness is used. It will also show that robustness is highly sensitive to specific combinations of circumstances, being influenced by a wide variety of strongly and complexly interacting factors. All this will be done under conditions that are entirely realistic to experimentation and statistical testing in the behavioral sciences.

METHOD

This study incorporates a fragment of the data and much of the methodology reported elsewhere (Bradley, 1980a, 1980b) in greater detail.

0090-5054/80/110333-04$00.65/0

Four populations, X, Y, x, and y, were used. All had the same mean. X and Y had a common standard deviation that was twice as large as the common standard deviation of x and y. Population X (see Bradley, 1980b) was L-shaped and closely resembled an empirical population of actual data (Bradley, 1976, 1977) encountered by the writer in a routine experiment, Population Y was bell-shaped (essentially normal), and Populations x and y were identical, respectively, to Populations X and Y, except for variance.

The following tests on means were investigated: the one-sample Z and t tests (denoted Z_1 and t_1) based upon N observations drawn from the X population; the two-correlated-sample t test (t_c) based upon N observations drawn from the X population and a correlated N observations from the Y population (see Bradley, 1980a); the two-independent-sample Z and t tests (Z_2 and t_2) based upon N_1 observations from one of the four populations (Population 1) and N_2 from either the same or a different one of the populations (Population 2), with N_1 and N_2 taking the relative values N and 3N, N and 2N, N and N, 2N and N, or 3N and N, respectively; and the multiindependent-sample F test (F) based upon three or four samples each of size N, drawn from one or two of the populations. Note that "N" is always the size of the smallest (or only) sample contributing to a test and therefore can serve as an index of absolute sample size.

For each of the following N values, 2, 4, 8, 16, 32, 64, 128, 256, 512, 1,024, an IBM 7090 computer, by means of pseudorandom sampling, obtained sampling distributions consisting of: (1) 50,000 values of the Z_1 test statistic and 50,000 values of the t_1 statistic, (2) 10,000 values of each of the other test statistics (Z_2, t_2, t_c, and F) under each different sampling situation investigated, except that for the Z_2, t_2, and t_c statistics when based upon equal-sized ($N_1 = N_2 = N$) samples from nonidentical populations, the N = 2 case was not investigated, so the smallest N value was 4.

For a given test statistic obtained under a set of conditions that are constant except for N, the sampling distributions obtained at different N values were independent in the sense that nonoverlapping and nonrecycling series of pseudorandom numbers were used to draw the samples contributing to them. Sampling distributions in other cases were not necessarily independent.

Many statistical checks were built into the program, such as sampling only from the Y population and testing the resulting distribution for predictable qualities. These checks showed that all was well.

For each sampling distribution of a test statistic, the proportion ρ of its values that occupied the rejection region for a test at the nominal (normal-theory) α level of significance was obtained. This was done for left-tailed, two-tailed, and right-tailed rejection regions corresponding to αs of .05, .01, and .001. Thus ρ is an estimate of the true probability of a Type I error (under whatever violations of assumptions have actually occurred), corresponding to an alleged probability α that would have been correct if no assumptions had been violated.

RESULTS

For those cases in which the tests are alleged to be robust (i.e., in which the expected value of ρ at infinite N is α for the Z and t tests, and sometimes not too far from α in the case of F), Tables 1, 2, and 3 give the N value at which ρ, in its final approach toward α, first entered a zone or belt extending from $\alpha - C$ to $\alpha + C$, where C represents a criterion for robustness. (The

Table 1
Changing Robustness of Z Tests With Increasing Absolute Sample Size: N Value at Which ρ First Entered the Zone $\alpha \pm C$ When Test Was Left-Tailed (L), Two-Tailed (T), and Right-Tailed (R)

Statistic	Pop.1	Pop.2	σ_1/σ_2	N_1	N_2	C = α/5 = .01 α = .05 L	T	R	C = α/2 α = .05 L	T	R	α = .01 L	T	R	α = .001 L	T	R
Z_1	X			N		128	4	64	32	2	4	256	32	128	—	128	—
				N	3N	64	4	32	16	2	4	64	16	32	256	32	512
				N	2N	32	4	4	2	2	4	2	16	32	2	32	32
Z_2	X	X	1	N	N	4	4	4	2	2	2	2	8	8	32	32	32
				2N	N	4	4	32	4	2	2	32	16	2	32	32	2
				3N	N	32	4	64	4	2	16	32	16	64	512	32	256
				N	3N	64	2	32	16	2	4	128	16	128	512	32	512
				N	2N	64	2	32	16	2	4	32	16	32	512	32	128
Z_2	X	Y	1	N	N	32	4†	4†	4†	4†	4†	16	4†	32	256	64	256
				2N	N	4	2	2	2	2	2	2	2	4	16	8	16
				3N	N	2	2	2	2	2	2	2	2	2	4	8	8
				N	3N	128	4	32	32	2	4	256	16	128	512	128	1024
				N	2N	128	4	32	32	2	4	128	32	256	512	32	1024
Z_2	X	x	2	N	N	64	4†	32	16	4†	4†	64	16	64	256	256	256
				2N	N	2	4	2	2	2	2	2	8	16	4	16	64
				3N	N	4	4	2	2	2	2	4	8	2	32	32	8
				N	3N	128	4	32	32	2	4	256	32	128	512	128	1024
				N	2N	128	4	32	32	2	4	128	16	128	512	128	1024
Z_2	X	y	2	N	N	128	4†	32	16	4†	4†	128	16	128	256	256	256
				2N	N	16	2	8	4	2	2	64	4	32	128	16	64
				3N	N	16	2	4	2	2	2	8	4	8	16	16	128
				N	3N	2	2	2	2	2	2	2	2	2	4	2	2
				N	2N	2	2	2	2	2	2	2	2	2	8	2	2
Z_2	Y	x	2	N	N	4†	4†	4†	4†	4†	4†	4†	4†	4†	32	16	8
				2N	N	2	2	4	2	2	2	16	4	4	16	16	16
				3N	N	8	2	32	2	2	2	32	4	16	256	32	64

Note–Heavier type indicates that in the neighborhood of the N value listed ρ was approaching α from above, lighter type from below. *†Smallest N value for which data were available in this case.*

Table 2
Changing Robustness of t Tests With Increasing Absolute Sample Size: N Value at Which ρ, in Its Final Approach to α, First Entered the Zone $\alpha \pm C$ When Test Was Left-Tailed (L), Two-Tailed (T), and Right-Tailed (R)

Statistic	Pop.1	Pop.2	$\frac{\sigma_1}{\sigma_2}$	N_1	N_2	$C = \alpha/5 = .01$			$C = \alpha/2$								
						$\alpha = .05$			$\alpha = .05$			$\alpha = .01$			$\alpha = .001$		
						L	T	R	L	T	R	L	T	R	L	T	R
t_1	X			N		—	256	1024	256	128	128	—	512	1024	—	—	—
t_2	X	X	1	N	3N	64	8	32	32	2	16	128	2	32*	256*	64*	32*
				N	2N	64	16	32*	16*	2	2	128*	8*	2	256*	16*	8
				N	N	16*	32*	16*	8*	16*	2	32*	64	32*	64	256	64
				2N	N	32*	16	64	2	2	16*	2	8*	128*	8	16*	256*
				3N	N	32	8	64	16	2	32	32*	2	128	32*	64*	256*
t_2	X	Y	1	N	3N	16	16	16	2	4	2	16*	8*	8*	32*	8	8
				N	2N	8*	8	64	2	2	8	2	2	64*	8	2	128*
				N	N	256	64	256	64	8	16	256	64	64*	512	128	512*
				2N	N	128	32	256	32	16	2	256	64	32*	1024	128	256
				3N	N	128	64	32	32	16	2	128	32	4	512	512	8
t_2	X	x	2	N	N	1024	128	512	128	64	64	512	256	1024	—	512	—
t_2	X	y	2	N	N	1024	128	512	256	64	64	1024	256	512	—	512	1024
t_2	Y	x	2	N	N	4†	16	16	4†	4†	4†	4†	16	16	8	32	32
t_2	Y	y	2	N	N	4†	4†	4†	4†	4†	4†	4†	4†	4†	8	16	4†
t_c	X	Y	1	N	N	64	16	64	8	4†	4†	64	16	4†	512	64	4†

*Note–Heavier type indicates that in the neighborhood of the N value listed ρ was approaching α from above, lighter type from below. *At some N value ≤ 4, ρ fell in the zone $\alpha \pm C$ but was not in its final approach to α and was outside the criterion zone at all higher N values below the value listed. †Smallest N value for which data were available in this case.*

Table 3
Changing Robustness of F Tests With Increasing Absolute Sample Size: N Value at Which ρ, in Its Final Approach to Its Expected Value at Infinite N, First Entered the Zone $\alpha \pm C$

Pop.1	Pop.2	Pop.3	Pop.4	$C = \alpha/5 = .01$	$C = \alpha/2$		
				$\alpha = .05$	$\alpha = .05$	$\alpha = .01$	$\alpha = .001$
X	X	X		64*	16*	16	64
X	X	X	X	16*	8*	8	8
X	Y	Y		8	4	16	64
X	Y	Y	Y	8	2	8	16
Y	X	X		8	4	32	64
Y	X	X	X	8	4	16	32
X	x	x		1024	64	—	—
X	x	x	x	—	64	—	—
x	X	X		32	16	32	—
x	X	X	X	16	2	16	32
X	y	y		1024	128	—	—
X	y	y	y	—	128	—	—
y	X	X		128	16	64	—
y	X	X	X	16	8	32	32
Y	x	x		128	8	—	—
Y	x	x	x	—	8	—	—
x	Y	Y		8	2	8	—
x	Y	Y	Y	16	2	16	128
Y	y	y		16	2	—	—
Y	y	y	y	—	2	—	—
y	Y	Y		4	2	8*	16*
y	Y	Y	Y	16*	2	16	16

*Note–Heavier type indicates that in the neighborhood of the N value listed ρ was approaching α from above, lighter type from below. *At N = 2, ρ was in the criterion zone $\alpha \pm C$ but was not in its final approach to its ultimate value; it departed from the criterion zone at N = 4 and reentered it at the N value listed.*

tabled N values are printed in heavier type for those cases in which ρ entered the criterion zone from above or in which ρ exceeded α at the smallest N value tested. Thus the heavier type generally signifies that ρ exceeds α at the tabled N value.) Two different criteria for robustness were used: a moderate criterion of $C = \alpha/5$, used only with $\alpha = .05$, so that the criterion was $.04 \leq \rho \leq .06$, and a very liberal criterion of $C = \alpha/2$, used for all three standard α levels, .05, .01, and .001, so that the criterion was $.5\alpha \leq \rho \leq 1.5\alpha$. This was the most liberal criterion the writer could bring himself to take seriously, since only twice as large a tolerable departure of ρ from α would include $\rho = 0$ and therefore categorize a test as robust whenever ρ fell below α. If α were the true probability of rejection, the probability of ρ falling within the criterion zone at any given N value would exceed .92 at $\alpha = .001$ and would vastly exceed .999 in all other cases. Had the tables listed the N value at which ρ finally entered (and subsequently remained within) a 95% tolerance band about α, then for each of the six tests at each of the α values listed, the medians of the table entries would have exceeded 100 and in 10 of the 18 cases, the "median" would have been a —, signifying that for over half the entries, ρ was not within the tolerance band at N = 1,024. Thus it clearly was not typical for the true probability of a Type I error to become statistically indistinguishable from α at small or moderate N values.

In the case of the Z tests, as N increased, ρ steadily approached α, except for fluctuations attributable to chance. So the N values given in Table 1 are always those at which ρ first entered the criterion zone, that is, first without any qualification. For the other tests, ρ sometimes took a circuitous route to its limiting

expected value, initially (i.e., at early N values) either departing farther from it or plunging through it, after which it then reversed direction and typically approached its limiting value monotonically, except for fluctuations attributable to chance. In some cases, therefore, ρ was inside (or on the other side of) the criterion zone at an early N value (usually 2, but occasionally a little higher), after which it departed from it in a circuitous path and either never reentered it or reentered it at a later N value, which is the N value given in the tables. In such cases, the first entry of the zone in ρ's final approach toward its limiting expected value is actually the second entry. These anomalous cases are indicated by an asterisk.

Data for cases in which $\sigma_1/\sigma_2 = 2$ and $N_1/N_2 = 3$, 2, 1/2, or 1/3 are presented only for the Z_2 test. For the t_2 test in these cases, the expected value of ρ at infinite N, obtained by formula or from graphs (see Bradley, 1968a, 1976), is always outside (and usually far outside) of the liberal criterion zone. It has frequently been claimed that t_2 is robust against nonnormality and heterogeneity when sample sizes are "not too unequal." Apparently, then, sizes in the ratio of 2 or 3 are too unequal.

DISCUSSION

The tables clearly show that (1) even when the conditions under which the tests are widely claimed to be robust are met, sample sizes in the hundreds or even thousands may not be sufficient to meet a very lax standard of robustness, and (2) effects are complex and highly qualified, being the result of numerous interacting factors. In about 25% of the cases investigated, the liberal criterion for robustness either was first met at an N value exceeding 100 or was never met, and for five of the six tests studied, there were a number of cases in which the criterion either was not met until N = 1,024 (2% of all cases investigated) or was never met (6%). In this connection, one should bear in mind that N is the size of the smallest sample involved in the test. There were cases in which the liberal criterion was not met even though the test statistic was based upon a total of 4,096 observations.

The results clearly show the complexity of the robustness phenomenon. Although the Z_2 test does not assume homogeneity of variance, its robustness against nonnormality is influenced (usually adversely) by heterogeneity of variance. When one population was skewed and the other was normal, Z_2's robustness tended to increase with the relative size of the sample drawn from the skewed population. When both populations were skewed, its robustness tended to improve as the ratio $(\sigma_1/\sqrt{N_1})/(\sigma_2/\sqrt{N_2})$ between the standard errors of the sample means drew closer to 1. And although Z_2 was highly robust at $\alpha = .05$ when the test was two-tailed, this was not necessarily the case when the test was left-tailed or when α was .01. Although the t_2 test is impressively robust against heterogeneity of variance alone when samples are of equal size, the additional violation of the normality assumption (by both populations) produced nonrobustness of much greater degree than would be expected from its robustness against nonnormality alone and heterogeneity alone (i.e., the effects of the two violations are interactive rather than additive). When sample sizes were equal, the robustness of the t_2 test tended to be greater when the population with the larger variance was normal. For the equal-sample F test, when all populations had the same variance, robustness tended to improve with increasing numbers of populations. When all populations had the same shape and all but one had the same variance, the F test was more robust when the exceptional population had the smaller variance. Finally, although the tables do not always show the effect, when the sampled populations are symmetric (e.g., Y and y) or when sample sizes are equal and populations are identical (X and X), robustness is worse for two-tailed than for one-tailed Z_2 and t_2 tests, whereas in most other cases robustness for a two-tailed test is either superior to or intermediate between the robustness of right-tailed and left-tailed tests at the same α level. These examples are illustrative rather than exhaustive. Thus a multitude of factors influence robustness, and interactions abound.

The results clearly show that even at very large sample sizes, the Z, t, and F tests are not unqualifiedly robust against nonnormality and that merely using samples of equal size does not necessarily make them robust against a combination of nonnormality and heterogeneity of variance. Furthermore, they show that degree of nonrobustness (i.e., roughly the relative departure of ρ from α) is highly dependent upon a variety of interacting factors. These include not only the test used, the assumptions violated, the degree of violation, and the criterion for robustness, but also nonviolative "testing" factors, such as the nominal value of α and the location of the rejection region, and "sampling" factors, such as the absolute and relative size of each sample and the particular shape and relative variance of the population from which that particular sample was drawn (i.e., the pairing of samples with populations). The multiplicity of influencing factors, the potential strength of their influence, and the complexity of their interactions with each other insure that a properly qualified claim of robustness will generally require a considerable amount of qualification—much more than is presently in vogue.

REFERENCES

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

BRADLEY, J. V. *Distribution-free statistical tests.* Englewood Cliffs, N.J: Prentice-Hall, 1968. (a)

BRADLEY, J. V. Studies in research methodology. *Dissertation Abstracts*, 1968, **28**, 4815B-4816B. (Monograph) (b)

BRADLEY, J. V. *Probability; decision; statistics.* Englewood Cliffs, N.J: Prentice-Hall, 1976.

BRADLEY, J. V. A common situation conducive to bizarre distribution shapes. *American Statistician*, 1977, **31**, 147-150.

BRADLEY, J. V. Robustness? *British Journal of Mathematical and Statistical Psychology*, 1978, **31**, 144-152.

BRADLEY, J. V. Nonrobustness in classical tests on means and variances: A large-scale sampling study. *Bulletin of the Psychonomic Society*, 1980, **15**, 275-278. (a)

BRADLEY, J. V. Nonrobustness in one-sample Z and t tests: A large-scale sampling study. *Bulletin of the Psychonomic Society*, 1980, **15**, 29-32. (b)

GLASS, G. V., & STANLEY, J. C. *Statistical methods in education and psychology.* Englewood Cliffs, N.J: Prentice-Hall, 1970.

SCHEFFÉ, H. *The analysis of variance.* New York: Wiley, 1959.

(Received for publication October 22, 1980.)