Bulletin of the Psychonomic Society
1980, Vol. 15 (4), 275-278

Nonrobustness in classical tests on means and variances: A large-scale sampling study

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

The robustness of the classical tests on means (Z, t, and F) and variances (chi square and F) was investigated by obtaining 30,000 (or, sometimes, 10,000 or 150,000) values of the test statistic under assumption-violating conditions and comparing the actual proportion of Type I errors with the proportion expected when all assumptions are met. The sampling and testing conditions investigated were: population shape (L-shape or bell-shape), relative population variance (1 or 4), sample size (8, 16, or 24), nominal significance level (.05, .01, or .001), and location of rejection region (left-tailed, right-tailed, or two-tailed). All tests on means were nonrobust under most of the investigated conditions, and one was nonrobust under all of them. Both tests on variances were extremely nonrobust under virtually all conditions. The worst nonrobustness of the one- and two-independent-sample Z and t tests rivaled or exceeded that of the notoriously nonrobust chi-square and F tests on variances.

Unqualified, or insufficiently qualified, claims of robustness for the classical Z, t, and F tests on means have become commonplace (Bradley, 1978), whereas the nonrobustness of the classical chi-square and F tests on variances is generally acknowledged (Scheffé, 1959). This article explores the robustness of all these tests under realistic conditions in terms of experimentation in the behavioral sciences. When the normality assumption is violated, this is accomplished by sampling from a population having an L-shape that is fairly common in certain areas of psychological and sociological research (Bradley, 1977); when the assumption of equal variances is violated, one population's standard deviation is only twice that of another. Samples are of reasonable absolute and relative size ($8 \leqslant N \leqslant 24$), and standard α levels and rejection regions are used. Yet under these perfectly realistic conditions, it will be seen that the robustness of tests on means depends heavily upon circumstances and that the Z and t tests are sometimes nonrobust to a degree that rivals or exceeds that of the admittedly nonrobust tests on variances.

METHOD

The Sampled Populations

Four different populations were constructed and labeled X, Y, x, and y. Population X was L-shaped and was constructed to resemble very closely a population of actual data (shown in Bradley, 1977, at the bottom of Figure B) obtained in a routine experiment (Bradley, 1980). Population Y was bell-shaped, exactly symmetric, and as nearly normal as its discreteness and limited range would permit. Populations X and Y had exactly the same mean and standard deviation and were similar in other ways. They are shown, and described in greater detail, in Bradley (1980).

By taking every unit in the X and Y populations and moving it halfway to the population mean, Populations x and y were obtained. They are therefore essentially the same as the X and Y populations, respectively, except that they have a common standard deviation that is exactly half as large as that of the X and Y populations.

The Tests Investigated

The robustness of eight different tests was investigated, always for the case in which the null hypothesis is true. The tests (and, in parentheses, the symbol used in this article to denote them) were: the one-sample Z test (Z_1) and the one-sample t test (t_1) of the hypothesis that the mean of the sampled population has a specified numerical value; the two-independent-sample Z test (Z_2), the two-independent-sample t test (t_2), the two-correlated-sample t test (t_c), and the multi-independent-sample F test (F) of the hypothesis that all sampled populations have the same unspecified mean; the one-sample chi-square test (χ^2_V) of the hypothesis that the variance of the sampled population has a specified numerical value; and the two-independent-sample F test (F_V) of the hypothesis that the two sampled populations have the same unspecified variance. All of these tests assume (in addition to the usual assumptions about random sampling) that every sampled population is normally distributed, and the t_2 and F tests also assume that all sampled populations have the same variance.

Sampling

Sample size was always 8 for the one-sample tests and for each of the three or four samples contributing to F. For the two-sample tests, sample sizes were 8 and 16, 8 and 24, or 16 and 16. The exact combinations of sampling and testing conditions under which the tests were investigated are given in the tables presented in the Results section.

Sampling and computations were performed by an IBM 7090 computer. For each test statistic under each different sampling situation (i.e., for each row of the tables), the computer obtained a sampling distribution of that test statistic under those sampling conditions. The sampling distributions for Z_1 and t_1 each contained 150,000 values of the test statistic, and that for χ^2_V contained 10,000 values. For all other test statistics, each sampling distribution was based upon 30,000 values of the test statistic.

Roughly speaking, independence of sampling existed within but not between sampling distributions. That is, nonoverlapping

0090-5054/80/040275-04$00.65/0

series of nonrecycling pseudorandom numbers were used to draw the different observations contributing to the sampling distribution of a given test statistic in a given sampling situation. However, the same or overlapping sets of pseudorandom numbers were used to obtain the sampling distributions of the same test statistic under different sampling conditions or of different test statistics under either the same or different sampling conditions.

Correlated samples were obtained for the t_c test by using the same set of pseudorandom numbers in the same way (Bradley, 1980) to draw the sample observations from each of the two sampled populations.

As a check on the "randomness" of sampling, the accuracy of programming, and the general validity of the study, "control" sampling distributions were obtained for each test statistic (except t_c) under each set of sample sizes for the case in which all of the samples come from the bell-shaped quasi-normal Population Y. Chi-square tests of fit to the appropriate normal-theory distribution of the test statistic showed that all was well.

RESULTS

For each test statistic under each different sampling situation, the computer obtained the proportion ρ of values of the test statistic that fell into a normal-theory rejection region corresponding to a nominal significance level of α. Thus ρ is the empirical probability (and therefore an estimate of the true probability) of a Type I error corresponding to an alleged probability α (that would have been the true probability if all assumptions had been met). This was done for left-tailed, right-tailed, and two-tailed rejection regions corresponding to α values of .05, .01, and .001.

Tables 1, 2, and 3 give the ρ values for the various tests under the various sets of circumstances that were investigated. The relatively robust ρ values lying in the interval from $2\alpha/3$ to $3\alpha/2$ are italicized; those lying outside of that interval but within the interval from $\alpha/3$ to 3α are printed in light ordinary type; those lying outside of that interval but within the interval from $\alpha/5$ to 5α are printed in medium type; and those lying outside of the latter interval, so that the larger of the two ratios ρ/α and α/ρ exceeds 5, are printed in heavy (boldface) type. This coding scheme refers to the exact values of ρ, so occasionally the tabled values, rounded to four decimal places, will appear to be misclassified. When ρ is exactly 0, it is written 0; when it becomes 0 only when rounded to the number of decimal places used, it is written .0000.

The most impressive case of robustness was that of

Table 1
Robustness of Z (and Other) Tests: Empirical Probability ρ of a Type I Error Corresponding to Nominal Probability α for Left-Tail (L), Two-Tail (T), and Right-Tail (R) Rejection Regions

Statistic	Pop. 1	Pop. 2	σ_1/σ_2	N_1	N_2	α = .05			α = .01			α = .001		
						L	T	R	L	T	R	L	T	R
z_1	X			8		**0**	*.0451*	*.0683*	**0**	.0207	.0295	**0**	**.0071**	**.0099**
				8	24	.0252	*.0486*	*.0646*	**.0030**	.0160	.0215	**.0001**	**.0043**	**.0059**
				8	16	*.0344*	*.0513*	*.0613*	.0058	*.0148*	.0189	.0005	**.0038**	**.0048**
z_2	X	X	1	16	16	*.0492*	*.0506*	*.0477*	*.0107*	*.0124*	*.0119*	.0019	.0018	*.0015*
				16	8	*.0613*	*.0513*	*.0344*	.0189	*.0148*	.0058	**.0048**	**.0038**	.0005
				24	8	*.0646*	*.0486*	.0252	.0215	.0160	**.0030**	**.0059**	**.0043**	**.0001**
				8	24	.0182	*.0441*	*.0652*	**.0004**	.0162	.0232	**0**	**.0045**	**.0065**
				8	16	.0253	*.0453*	*.0619*	**.0011**	*.0141*	.0221	**0**	**.0040**	**.0060**
z_2	X	Y	1	16	16	*.0384*	*.0481*	*.0560*	.0051	*.0108*	.0162	.0004	.0018	.0024
				16	8	*.0452*	*.0487*	*.0543*	*.0072*	*.0100*	*.0127*	.0004	*.0012*	.0020
				24	8	*.0468*	*.0506*	*.0528*	*.0084*	*.0101*	*.0124*	*.0008*	*.0010*	*.0015*
				8	24	**.0053**	*.0440*	*.0690*	**0**	.0178	.0256	**0**	**.0058**	**.0081**
				8	16	**.0106**	*.0455*	*.0678*	**.0002**	.0168	.0252	**0**	**.0056**	**.0076**
z_2	X	x	2	16	16	.0312	*.0478*	*.0608*	**.0031**	*.0142*	.0196	**.0002**	.0028	**.0047**
				16	8	*.0434*	*.0526*	*.0540*	*.0083*	*.0142*	.0160	*.0012*	.0027	**.0031**
				24	8	*.0497*	*.0538*	*.0479*	*.0126*	*.0140*	*.0115*	.0023	.0027	.0017
				8	24	**.0023**	*.0428*	*.0700*	**0**	.0180	.0261	**0**	**.0057**	**.0080**
				8	16	**.0050**	*.0432*	*.0678*	**0**	.0174	.0252	**0**	**.0057**	**.0079**
z_2	X	y	2	16	16	.0264	*.0450*	*.0624*	**.0011**	*.0139*	.0216	**0**	**.0033**	**.0048**
				16	8	*.0334*	*.0483*	*.0587*	**.0028**	*.0130*	.0193	**.0000**	.0026	**.0040**
				24	8	*.0395*	*.0502*	*.0570*	.0054	*.0113*	.0160	**.0001**	.0022	**.0032**
				8	24	*.0517*	*.0495*	*.0498*	*.0107*	*.0096*	*.0088*	*.0012*	*.0011*	*.0011*
				8	16	*.0498*	*.0494*	*.0486*	*.0105*	*.0100*	*.0085*	*.0012*	*.0010*	*.0009*
z_2	Y	x	2	16	16	*.0513*	*.0494*	*.0473*	*.0117*	*.0101*	*.0081*	.0018	*.0014*	.0004
				16	8	*.0529*	*.0490*	*.0407*	.0157	*.0121*	.0063	.0025	.0017	**.0003**
				24	8	*.0561*	*.0481*	*.0384*	.0172	*.0131*	.0043	**.0037**	.0025	**.0002**
F_v	X	X	1	16	16	**.2725**	**.4809**	**.2687**	**.2134**	**.3944**	**.2116**	**.1738**	**.3336**	**.1734**
F_v	X	Y	1	16	16	**.2788**	**.3256**	.1136	**.2244**	**.2416**	**.0440**	**.1965**	**.1996**	**.0110**
χ^2_v	X			8		**.4383**	**.5543**	.1608	**.4313**	**.4919**	**.0819**	**.4290**	**.4555**	**.0364**

0 ≤ heavy < $\alpha/5$ ≤ medium < $\alpha/3$ ≤ light < $2\alpha/3$ ≤ *italics* ≤ $3\alpha/2$ < light ≤ 3α < medium ≤ 5α < heavy ≤ 1.00

Table 2
Robustness of t Tests: Empirical Probability ρ of a Type I Error Corresponding to Nominal Probability α for Left-Tail (L), Two-Tail (T), and Right-Tail (R) Rejection Regions

Statistic	Pop. 1	Pop. 2	σ_1/σ_2	N_1	N_2	α = .05			α = .01			α = .001		
						L	T	R	L	T	R	L	T	R
t_1	X			8		.4283	.4277	.0044	.4198	.4045	.0001	.3129	.2501	.0000
				8	24	.0038	*.0414*	*.0737*	.0002	*.0118*	.0216	.0000	.0017	.0033
				8	16	.0101	*.0350*	*.0638*	.0009	.0054	*.0101*	.0001	.0004	*.0008*
t_2	X	X	1	16	16	*.0445*	.0279	*.0429*	.0026	.0016	.0026	.0001	.0001	.0001
				16	8	*.0638*	*.0350*	.0101	*.0101*	.0054	.0009	*.0008*	.0004	.0001
				24	8	*.0737*	*.0411*	.0038	.0216	*.0118*	.0002	.0033	.0017	.0000
				8	24	*.0441*	*.0336*	*.0377*	.0043	.0049	*.0067*	.0002	.0003	.0006
				8	16	*.0621*	*.0432*	.0293	*.0127*	*.0073*	.0035	*.0011*	.0006	.0001
t_2	X	Y	1	16	16	.0833	*.0607*	.0252	.0276	.0190	.0021	.0066	.0042	.0000
				16	8	.1055	.0853	.0293	.0486	.0375	.0041	.0191	.0148	.0005
				24	8	.0978	.0831	*.0351*	.0458	.0368	.0053	.0182	.0151	.0006
				8	24	.2803	.1858	.0756	.0557	.0474	.0188	.0215	.0192	.0020
				8	16	.2849	.1796	*.0410*	.0800	.0615	.0057	.0377	.0322	.0005
t_2	X	x	2	16	16	.1885	.1665	.0138	.1140	.0730	.0004	.0250	.0172	0
				16	8	.1804	.1513	.0007	.1029	.0702	0	.0222	.0131	0
				24	8	.1007	.0824	.0002	.0640	.0494	0	.0214	.0131	0
				8	24	.2288	.1783	*.0651*	.0710	.0469	*.0119*	.0081	.0050	.0016
				8	16	.2192	.1636	*.0376*	.0812	.0507	.0046	.0153	.0085	.0003
t_2	X	y	2	16	16	.1550	.1262	.0136	.0904	.0694	.0004	.0358	.0261	0
				16	8	.1389	.1123	.0025	.0861	.0708	.0001	.0435	.0348	.0001
				24	8	.0895	*.0693*	.0012	.0523	.0436	.0001	.0301	.0255	.0000
				8	24	.1141	.1711	.1341	.0431	.0721	.0613	.0108	.0239	.0226
				8	16	.0877	.1266	.1098	.0261	.0463	.0448	.0051	.0141	.0150
t_2	Y	x	2	16	16	*.0441*	*.0520*	*.0582*	*.0079*	*.0123*	*.0147*	.0006	.0016	.0024
				16	8	.0214	.0183	.0239	.0021	.0018	.0029	.0000	.0001	.0002
				24	8	.0156	.0080	.0101	.0017	.0007	.0007	.0001	.0001	0
				8	24	.1180	.1566	.1168	.0460	.0612	.0451	.0115	.0154	.0117
				8	16	.0945	.1141	.0928	.0294	.0371	.0312	.0056	.0069	.0058
t_2	Y	y	2	16	16	*.0508*	*.0516*	*.0510*	*.0105*	*.0110*	*.0102*	*.0013*	*.0015*	*.0012*
				16	8	.0246	.0198	.0232	.0028	.0022	.0029	.0002	.0001	.0001
				24	8	.0146	.0091	.0133	.0012	.0009	.0012	.0001	.0001	.0000
t_c	X	Y	1	16	16	*.0645*	*.0527*	*.0375*	.0182	*.0139*	.0057	.0037	.0023	.0005

0 ≤ heavy < $\alpha/5$ ≤ medium < $\alpha/3$ ≤ light < $2\alpha/3$ ≤ *italics* ≤ $3\alpha/2$ < light ≤ 3α < medium ≤ 5α < heavy ≤ 1.00

the two-tailed Z_1 and Z_2 tests at $\alpha = .05$, in which ρ usually fell within a 10% deviation from α and always within 20%. The t_c test was also reasonably robust at $\alpha = .05$; and when sample sizes were equal and only the assumption of equal variances was violated, the t_2 test was robust at all α values. However, it is clear from the tables that there are few easily and succinctly generalizable patterns of robustness.

Extreme departures of ρ from α were common for most of the tests on means. Often ρ was many times greater than α (more than 312 times as great in the worst case) or α was many times greater than ρ (infinitely greater in the 20 cases in which ρ was 0). Yet even at the relatively innocuous α level of .05, multiples exceeding 3 occurred for all the table entries for t_1, for a third of the entries for Z_1 and t_2, and for some of the entries for Z_2. And only for the t_c test did all of the table entries at $\alpha = .05$ fall in the interval from $2\alpha/3$ to $3\alpha/2$. The latter is an ultraliberal criterion for robustness, since a lower boundary only three times as far from α would include $\rho = 0$ so that all ρ values below α would be called robust.

Perhaps the most startling result of this study, however, is that there were numerous instances in which the nonrobustness of the reputedly robust Z and t tests on means rivaled or exceeded that of the notoriously nonrobust χ^2_v or F_v tests on variances. If we measure nonrobustness by the larger of the two ratios ρ/α and α/ρ and make comparisons at the same α value, we find that at all α values the t_1 test is about as nonrobust and the left-tailed Z_1 test is more nonrobust than the χ^2_v test at the same sample size; both tests are usually more nonrobust than the F_v test based on larger samples. Furthermore, there are 22 cases for t_2 and 15 for Z_2 in which their nonrobustness exceeds that of the median table entry for F_v tests at the same α value, and 20%-40% of these cases occur at the most robustness-conducive α value of .05. Finally, the nonrobustness of about 1/2 the table entries for t_2, 1/10 for Z_2, and 1/7 for F exceeds that of the smallest table entry for F_v at the

Table 3
Robustness of the Multisample F Test When N Observations Are Drawn From Each of Three or Four Sampled Populations

				Empirical Probability ρ of a Type I Error Corresponding to Nominal Probability α When N = 8		
Pop.1	Pop.2	Pop.3	Pop.4	α = .05	α = .01	α = .001
X	X	X		.0248	.0036	.0003
X	X	X	X	.0314	.0052	.0006
X	Y	Y		*.0604*	.0156	.0026
X	Y	Y	Y	*.0562*	*.0138*	.0025
Y	X	X		*.0563*	.0216	**.0083**
Y	X	X	X	*.0536*	.0185	**.0063**
X	x	x		.1275	**.0480**	**.0210**
X	x	x	x	.0988	**.0318**	**.0115**
x	X	X		.1093	**.0349**	**.0078**
x	X	X	X	*.0714*	.0204	**.0032**
X	y	y		.1281	**.0451**	**.0084**
X	y	y	y	.1082	**.0324**	**.0057**
y	X	X		.0916	**.0481**	**.0206**
y	X	X	X	*.0682*	.0292	**.0129**
Y	x	x		*.0674*	.0236	**.0061**
Y	x	x	x	*.0727*	.0267	**.0083**
x	Y	Y		*.0597*	.0158	.0027
x	Y	Y	Y	*.0601*	.0161	.0024
Y	y	y		*.0652*	.0184	.0026
Y	y	y	y	*.0709*	.0222	**.0040**
y	Y	Y		*.0573*	.0151	.0021
y	Y	Y	Y	*.0579*	.0158	.0021

Larger of ρ/α and α/ρ is: ≤ *1.5*, 1.5 to 3, 3 to 5, > 5

same α value, and always at least 1/5 of these cases occur at α = .05. Thus there were occasions when each of the tests on means except t_c was nonrobust to a degree encountered in the tests on variances, and these occasions were quite frequent for t_1, t_2, and Z_1.

DISCUSSION

Objections are sometimes raised against using the same relative criterion for robustness at all three α levels rather than using increasingly lax criteria at increasingly remote rejection regions. However, if one has properly chosen his α level as the maximum risk of Type I error that he is willing to countenance (after due consideration of power requirements and tradeoffs between risks of Type I and Type II errors), it is about as disastrous for that risk to be double what he thought when α = .001 as when α = .05.

It is sometimes claimed that it is less serious for ρ to fall below α than to fall above it because in the former case the actual probability of a Type I error is on the "safe" or "conservative" side of the nominal probability α. However, this is no cause for complacency, since, in that case, the power of the test is reduced accordingly and the probability of a Type II error must therefore be on the "unsafe" or "radical" side of β. Therefore, ironically, a nonparametric test using the same nominal α level and the same sample sizes may have considerably more power than the test actually selected because of the robustness and superior efficiency claimed for it.

As mentioned earlier, none of the conditions investigated in this study were unrealistic to actual experimentation in the behavioral sciences. On the contrary, most of the investigated conditions were quite mild. Consequently, the present study cannot be regarded as having explored the outer reaches of nonrobustness—neither the mathematical boundaries nor the extremes that might actually be encountered in practice—nor as having remotely approached such an exploration. Yet the majority of the ρ values found for each of the tests failed to meet a rather lax criterion of robustness. Worse, what robustness was found occurred sporadically and was highly peculiar to specific sampling conditions. Consequently, it is seldom possible to formulate an adequately, but not profusely, qualified statement of any set of conditions under which robustness is invariably attained by a given test, not only in general but even within the limited range of (usually mild) conditions investigated in the present study. This resistance of robustness to succinct generalization greatly weakens the usefulness of the term.

REFERENCES

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 one-sample Z and t tests: A large-scale sampling study. *Bulletin of the Psychonomic Society*, 1980, **15**, 29-32.

SCHEFFÉ, H. *The analysis of variance.* New York: Wiley, 1959. Pp. 336-337, 360.

(Received for publication February 18, 1980.)