Bulletin of the Psychonomic Society
1980, Vol. 15 (1), 29-32

Nonrobustness in one-sample Z and t tests: A large-scale sampling study

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

For each of the N-values 2, 4, 8, 16, 32, 64, 128, 256, 512, and 1,024, 50,000 samples of size N were drawn from an L-shaped population, and for each sample the Z and t statistics were calculated. The resulting distributions of 50,000 Z or t values at each sample size were then used to study the robustness of left-tailed, right-tailed, and two-tailed Z and t tests at α levels of .05, .01, and .001 (and, for Z only, .0001). The actually obtained proportion, ρ, of Type I errors was often far greater or far smaller than the nominal proportion, α. Furthermore, although α is the expected value of ρ at infinite N, no N-value below 512 ever brought the deviation of ρ from α to within 10% of α for any t tests or one-tailed Z tests.

Some very rash claims have been made about the robustness of the one-sample Z and t tests (see Bradley, 1978). However, there are some perfectly realistic and common experimental circumstances (Bradley, 1977) under which the tests are quite nonrobust. One such situation is that in which samples are drawn from a long-tailed L-shaped population. This article is concerned with that situation. Incomplete portions of the results have appeared elsewhere (Bradley, 1976) as tersely explained graphs. The purpose of this article is to present the complete numerical data and methodology.

METHOD

Sampled Populations

A single experimental condition was selected from the many treatment conditions that had been investigated in a routine experiment. A single subject was then given 2,520 trials under this single condition. The subject's task was to release a telegraph key that he had been depressing with his right hand, then reach up with the same hand and operate a pushbutton, whose proper operation provided feedback by extinguishing a light. The locations of both the telegraph key and the pushbutton were rigidly fixed, so the distance between them was held constant. A time clock measured, in hundredths of a second, the time elapsed from the moment the telegraph key was released until the moment the pushbutton was successfully operated, turning out the light. (So the time measured does not include reaction time, which must end before the release of the telegraph key.)

The frequency distribution of the 2,520 time scores for this subject was L-shaped and is shown in another article (Bradley, 1977, bottom of Figure B or top of Figure C). Using this distribution as a model, a smoother and more regular frequency distribution of 100,000 integer-valued scores was constructed and called the X population (see Figure 1). The X population closely resembles the original distribution of time scores in all its essential characteristics. However, its skewness (3.18) and kurtosis (13.85) are both somewhat smaller, and closer to those (0 and 3) of a normal distribution, than are those (3.42 and 17.29) of the original distribution of time scores. Consequently, using the X population as the sampled population will be conservatively more conducive to robustness than if the original distribution of time scores had been used.

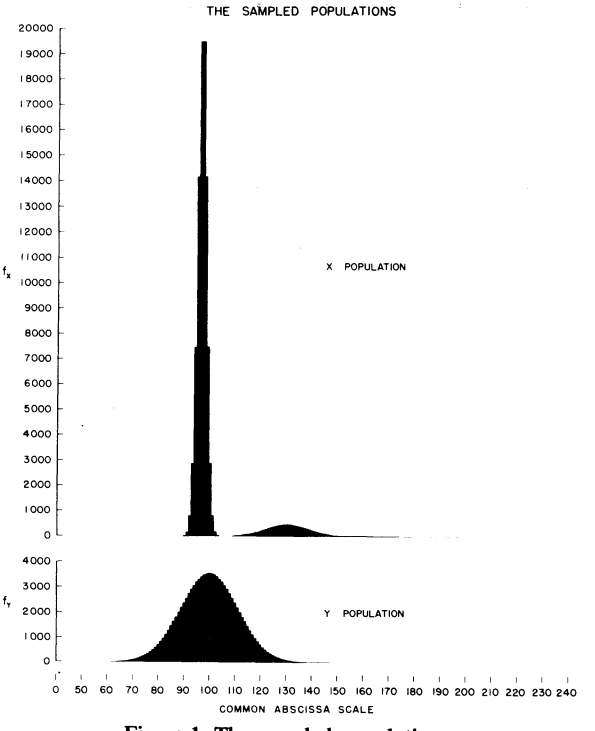

Figure 1. The sampled populations.

A second frequency distribution of 100,000 integer-valued scores was then constructed, having exactly the same mean and variance as the X population but being perfectly symmetric and as bell-shaped (Gaussian) as was possible to obtain with this finite number of integer-valued units. This quasinormal distribution, which we shall call the Y population (see Figure 1), will be used to produce "control" samples whose Z and t statistics should be highly robust if all operations and calculations have been properly performed (and if the discreteness and finite range of the population have no appreciable effect). Since most of these operations and calculations (such as the method of sampling) will be the same for the X population, robustness of a test when sampling from the Y population will tend to validate whatever results are obtained when sampling under otherwise identical conditions from the X population.

0090-5054/80/010029-04$00.65/0

Sampling

All sampling and calculations were performed by an IBM 7090 computer. Pseudorandom numbers ranging from 1 to 100,000 were generated by a formula whose repetition cycle was long enough to produce all of the required data without any overlap in the series of pseudorandom numbers due to recycling. In order to make the X and Y samples maximally comparable, a sample of N X-observations and its counterpart sample of N Y-observations were drawn by using the same set of N consecutive pseudorandom numbers to identify and draw observations from the two populations in exactly the same way. Each pseudorandom number R (where $1 \leq R \leq 100{,}000$) was used to draw (with replacement) the Rth, in order of increasing value, of the 100,000 units in the X population for the X sample and the Rth, in order of increasing value, of the 100,000 units in the Y population for the Y sample. Thus for every observation in the X sample, there was a counterpart observation in the corresponding Y sample; the counterpart X and Y observations were values having exactly equal ordinates in the cumulative frequency distributions of the X and Y populations, respectively.

Using this method, for each of the N-values 2, 4, 8, 16, 32, 64, 128, 256, 512, and 1,024, the computer drew 50,000 X samples and 50,000 counterpart Y samples, each consisting of N observations. For each sample of size N drawn from the X population, the computer calculated:

$$Z_X = \frac{\bar{X} - \mu}{\sqrt{\sigma^2/N}}$$

and

$$t_X = \frac{\bar{X} - \mu}{\sqrt{\Sigma(X - \bar{X})^2/N(N-1)}}$$

and calculated for the N observations drawn from the counterpart Y sample:

$$Z_Y = \frac{\bar{Y} - \mu}{\sqrt{\sigma^2/N}}$$

and

$$t_Y = \frac{\bar{Y} - \mu}{\sqrt{\Sigma(Y - \bar{Y})^2/N(N-1)}},$$

substituting the common numerical mean of the X and Y populations for μ and the common numerical variance of the two populations for σ^2. Since the correct numerical value was substituted for μ, the four statistics can all be regarded as test statistics used to test a true null hypothesis ($H_0: \mu = \mu_0$) about the value of the population mean. We shall therefore be investigating robustness of Type I error (as contrasted with robustness of power).

Thus, for each of the four statistics listed above, and at each of the 10 sample sizes, the computer obtained a sampling distribution of 50,000 values of that statistic at that sample size. From each of these sampling distributions, the computer tabulated the proportion ρ of statistic-values falling into a rejection region that would be expected to contain a proportion α of such values if all the test's assumptions, including normality, had been met. This was done for left-, right-, and two-tailed rejection regions, and for α values of .05, .01, .001, and, for Z only, .0001. So α is the alleged significance level, that is, the probability of a Type I error had all assumptions been met; ρ is the empirical estimate of the true significance level, that is, of the actual probability of a Type I error under the actually prevailing conditions of sampling from the L-shaped or bell-shaped population and using a normal-theory rejection region of nominal size α.

RESULTS AND DISCUSSION

The empirical probability ρ of a Type I error corresponding to an alleged probability α, which would have been correct if the normality assumption had been met, is given for various α values, rejection region locations, sample sizes, and sampled populations, in Table 1 for the Z test and in Table 2 for the t test. For example, the normal-theory critical value for a right-tailed Z test at $\alpha = .05$ is 1.644853637. At N = 2, some 5,417 of the 50,000 Z_X values, or a proportion .10834 of them, exceeded this critical value, and this is the cell entry ρ for Z_X, N = 2, $\alpha = .05$, and right-tail rejection region.

The ρ values for the control statistics Z_Y and t_Y corresponded closely to α, their expected value under normal theory, except at the smallest N-values, at which some distortion occurred due to the discreteness of the sampled population. This distortion was greatest for t_Y at N = 2, at which there was an appreciable probability that all observations in a sample would have the same value, thereby causing the denominator of the t statistic to be zero and the resulting t value to be either plus or minus infinity.

Clearly, the robustness of the one-sample Z and t tests cannot be taken for granted. For t_X and Z_X, ρ was often many times greater than α, and α was often many times greater than ρ (infinitely greater in the numerous cases in which ρ was zero). Furthermore, many of these large multiples occurred at sample sizes typical of actual psychological experimentation. For t_X at α values of .05, .01, and .001, respectively, the ratio ρ/α ranged from .034 to 11.847, .004 to 41.860, and 0 to 312.70. For Z_X the analogous figures are 0 to 2.167, 0 to 3.812, and 0 to 16.42.

One of the commonest bases for the claim that Z and t are robust is the mathematically demonstrable fact that under nonnormality the true probability of a Type I error must equal α when N becomes infinite. This implies that the true probability must approach α as an asymptote, and this implication has encouraged many to claim robustness at very moderate N-values. Yet sample sizes far in excess of those generally employed in psychological research were usually required to bring the deviation of ρ from α to within 10% of α when sampling from the L-shaped population. Indeed, this degree of closeness occurred for t_X in only 2 of the 90 entries in the table: only when $N \geq 512$ and the test was two-tailed at a nominal α of .05. In the case of Z_X, it occurred 14 times—always for a two-tailed rejection region except when $\alpha = .05$ and $N \geq 512$. In fact, for t_X the deviation of ρ from α never fell within even 50% of α, at any α value or for any rejection region, until N = 128. For $\alpha = .05$, it first fell that close at N = 128 (but only for two-tailed and right-tailed tests);

Table 1
Robustness of the One-Sample Z Test When Sampling from the L-Shaped Population (X) and from the Bell-Shaped Population (Y)

N	Population	$\alpha = .05$			$\alpha = .01$			$\alpha = .001$			$\alpha = .0001$	
		L	T	R	L	T	R	L	T	R	L	R
2	X	0	.06718	.10834	0	.02588	.03812	0	.01430	.01642	0	.01104
	Y	*.04538*	*.04756*	*.04850*	*.01028*	*.01100*	*.01110*	*.00126*	*.00110*	*.00102*	*.00012*	*.00016*
4	X	0	.04998	.06418	0	.02658	.03562	0	.00884	.01156	0	.00502
	Y	*.05146*	*.05296*	*.05122*	*.01002*	*.01038*	*.00960*	*.00106*	*.00068*	*.00084*	*.00004*	*.00010*
8	X	0	.04492	.06902	0	.02030	.02912	0	.00664	.00984	0	.00352
	Y	*.04964*	*.04962*	*.04814*	*.01098*	*.00994*	*.00940*	*.00084*	*.00098*	*.00098*	*.00014*	*.00016*
16	X	.00110	.04348	.06792	0	.01630	.02374	0	.00474	.00694	0	.00230
	Y	*.05062*	*.05094*	*.05082*	*.00962*	*.00996*	*.01026*	*.00092*	*.00092*	*.00092*	*.00010*	*.00008*
32	X	.02942	.04026	.06144	0	.01192	.01944	0	.00304	.00462	0	.00122
	Y	*.05104*	*.04950*	*.04886*	*.01044*	*.01004*	*.00912*	*.00128*	*.00102*	*.00086*	*.00012*	*.00002*
64	X	.03534	.04686	.06000	.00226	.01128	.01756	0	.00242	.00370	0	.00074
	Y	*.04914*	*.04960*	*.04976*	*.01032*	*.00996*	*.01012*	*.00094*	*.00094*	*.00110*	*.00006*	*.00012*
128	X	.04036	.04692	.05754	.00396	.00988	.01500	.00012	.00144	.00258	0	.00030
	Y	*.04828*	*.04868*	*.05052*	*.00902*	*.00992*	*.01030*	*.00100*	*.00094*	*.00084*	*.00012*	*.00004*
256	X	.04062	.04914	.05624	.00560	.01044	.01440	.00026	.00148	.00238	0	.00046
	Y	*.04918*	*.04988*	*.05016*	*.00964*	*.00960*	*.01016*	*.00084*	*.00102*	*.00104*	*.00008*	*.00010*
512	X	.04500	.04862	.05472	.00762	.00980	.01214	.00040	.00112	.00170	.00002	.00020
	Y	*.04950*	*.04916*	*.04862*	*.00934*	*.00950*	*.00930*	*.00102*	*.00110*	*.00106*	*.00010*	*.00016*
1024	X	.04780	.05048	.05270	.00840	.01000	.01194	.00048	.00088	.00164	0	.00024
	Y	*.04872*	*.04822*	*.04962*	*.00976*	*.00958*	*.00956*	*.00104*	*.00090*	*.00094*	*.00010*	*.00004*

Note–Cell entries give proportion of empirical sampling distributions of 50,000 Zs falling in normal-theory left-tail (L), two-tail (T), or right-tail (R) rejection region of nominal size α.

Table 2
Robustness of the One-Sample t Test When Sampling from the L-Shaped Population (X) and from the Bell-Shaped Population (Y)

N	Population	$\alpha = .05$			$\alpha = .01$			$\alpha = .001$		
		L	T	R	L	T	R	L	T	R
2	X	.28580	.13710	.00588	.12962	.13032	.00128	.12962	.13002	.00040
	Y	*.05158*	*.05516*	*.05190*	*.01414*	*.02580*	*.01386*	*.01300*	*.02580*	*.01280*
4	X	.59234	.48872	.00170	.30834	.18618	.00004	.04898	.02928	0
	Y	*.04998*	*.05042*	*.05030*	*.01002*	*.01038*	*.01008*	*.00088*	*.00102*	*.00110*
8	X	.42700	.42618	.00382	.41860	.40306	.00004	.31270	.25024	0
	Y	*.05024*	*.04898*	*.04944*	*.01046*	*.00976*	*.00898*	*.00116*	*.00110*	*.00072*
16	X	.20402	.19706	.00802	.19084	.18928	.00014	.18766	.18722	0
	Y	*.05094*	*.04992*	*.05044*	*.01004*	*.00990*	*.00996*	*.00094*	*.00094*	*.00088*
32	X	.14522	.11646	.01204	.08444	.06952	.00042	.05106	.04688	0
	Y	*.05070*	*.05004*	*.04944*	*.01008*	*.01002*	*.00932*	*.00120*	*.00096*	*.00080*
64	X	.10782	.08322	.01978	.05268	.04086	.00112	.02404	.01978	0
	Y	*.04948*	*.04974*	*.05010*	*.00986*	*.00984*	*.00974*	*.00112*	*.00094*	*.00102*
128	X	.08724	.06648	.02608	.03508	.02444	.00194	.01118	.00856	.00002
	Y	*.04824*	*.04830*	*.05056*	*.00916*	*.00946*	*.01056*	*.00088*	*.00080*	*.00080*
256	X	.07258	.05636	.03278	.02404	.01686	.00352	.00632	.00418	.00014
	Y	*.04902*	*.05000*	*.05022*	*.00938*	*.00970*	*.01004*	*.00092*	*.00090*	*.00098*
512	X	.06678	.05312	.03614	.01992	.01418	.00444	.00432	.00260	.00014
	Y	*.04938*	*.04882*	*.04896*	*.00932*	*.00938*	*.00948*	*.00110*	*.00102*	*.00100*
1024	X	.06308	.05290	.04042	.01682	.01276	.00598	.00280	.00180	.00026
	Y	*.04894*	*.04792*	*.04974*	*.00974*	*.00960*	*.00942*	*.00108*	*.00098*	*.00086*

Note–Cell entries give proportion of empirical sampling distributions of 50,000 ts falling in normal-theory left-tail (L), two-tail (T), or right-tail (R) rejection region of nominal size α.

for $\alpha = .01$, the criterion was first met at $N = 512$ (but only for two-tailed tests); and for $\alpha = .001$, it never fell that close at any $N \leqslant 1{,}024$.

It appears, then, that, when based upon nonhuge samples from the L-shaped X population, whatever robustness the Z test has is highly dependent upon specific circumstances (tailedness and α level as well as N), and that the t test is nonrobust under all circumstances.

REFERENCES

BRADLEY, J. V. *Probability; decision; statistics.* Englewood Cliffs, N.J: Prentice-Hall, 1976. Pp. 390-405, 464.

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.

(Received for publication October 22, 1979.)