Sensitivity of Tests of Significance: A Problem Posed

R. C. GEARY

THIS paper contains a preliminary examination of the effect of using the dichotomy (0, 1) instead of actual measurement in the study of relationship. First, the simplest case possible is dealt with, in the hope of inspiring, or provoking, more thorough treatment by others.

Of course, dichotomy or other ordinal treatment has to be used when measurement is inconceivable, e.g. when the variable is sex or religion. It is very familiar in dummy variable practice in regression, e.g. in investigating relationship between time series when certain major events (e.g. a dock strike) are known *a priori* to have affected one or more of the data. It is thought that a comparison of statistical functions pertaining to relationship, when cardinal values are possible, will shed light on the meaning of these functions when ordinal values only are available.

Using two other examples we reconsider the customary χ^2 treatment of 2×2 tables. We end up with a very general query bearing on stochastic decision-making, arising out of our simple examples, for consideration by our colleagues.

Two Tests of Relationship

Let (X_i', Y_i'), $i = 1, 2 \ldots, N$, be measures of a pair of variables in a population of N. Coefficient of correlation is ρ which, without loss of generality, may be assumed $\geqslant 0$.

With M_x and M_y as the respective medians, new pairs (X_i, Y_i) are derived by assigning the value 1 to X_i when $X_i \geqslant M_x$; similarly for Y_i; otherwise values 0. Hence (X_i, Y_i) are now in 4 classes, (1, 1), (1, 0), (0, 1) (0, 0). When X_i' and Y_i' are independent of one another, and N indefinitely large, probability of occurrence of similars, (1, 1) and (0, 0), is 1/2. Hence degree of relationship between the X_i and Y_i can be adjudged, on a stochastic scale, by the amount actual probability κ exceeds 1/2.

The Bivariable Normal Case

We now assume that our original pair (X', Y') are continuous and distributed normally, i.e. with probability element—

$$(1) \qquad f(x, y) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2(1-\rho^2)}(x^2-2\rho xy+y^2)\right]$$

where (x, y) is the standardised version of (X', Y'), i.e. with means (0, 0), variances (1, 1) and correlation ρ. Population-wise, the unitary proportion of similars (1, 1) and (0, 0) will be given by—

$$(2) \qquad \kappa = \int_0^\infty \int_0^\infty [f(x, y)+f(-x, -y)]\, dxdy$$

But, from (1), $f(x, y) = f(-x, -y)$, so that—

$$\kappa = 2\int_0^\infty \int_0^\infty f(x, y)\, dxdy$$

$$(3) \qquad = \frac{1}{2}+\frac{1}{\pi}\sin^{-1}\rho,$$

as is well known. Note from (3) that if $\rho = 0$, $\kappa = 1/2$ and when $\rho = 1$, $\kappa = 1/2+(\sin^{-1} 1)/\pi = 1$, as they should.

Efficiency of Tests Compared

The main object of this note is to compare the efficiency of measures of relationship r and k, random sampling estimates of ρ and κ, for decision as to relationship (or not) between X and Y. In advance we suspect that r is superior since it uses more information (a hazardous criterion, as statisticians familiar with the rectangular distribution know). We shall find that this is so, and also show how much better.

Essentially we use the average critical value (ACV) approach,[1] a simplified substitute for full power function treatment.[2]

Following is a quotation, defining ACV, from the basic paper:—

> Suppose we have a test function $c(x_1, x_2, \ldots, x_n)$ where $x_1, x_2, \ldots, x_n$ are the measures of n random drawings from a population $f(x, \mu)$ of defined form but with a parameter μ (not necessarily the population mean), the value of which is at present undetermined. We wish to decide from the sample, using the test c, whether in the population the value of μ could plausibly be taken as zero (hypothesis H_0) or whether μ has probably some value greater than zero (hypothesis H_1). Clearly if μ is very small (without defining "smallness"), no test will be sensitive enough to yield an answer (when n is given and finite). It is proposed to reject the hypothesis H_0 that in the population $\mu = 0$ and accept the hypothesis H_1 when the value found for c in the particular sample is greater than, say, its 0·95 probability point λ, on the H_0 hypothesis. If $\mu > 0$ we are naturally interested in the values of c which are "near" λ, some greater and some less. Therefore set
>
> $$Ec(\mu) = \lambda, \tag{1.1}$$
>
> where E is the "expected" value, or the mean of an indefinitely large number of sample values. Assuming that $Ec(\mu)$ varies monotonically with μ, (1.1) is solved for μ by an identifiably unique value $\mu = M$, the ACV, so that
>
> $$M = \phi(\lambda). \tag{1.2}$$
>
> Now if there are two tests c_1 and c_2 with 0·95 probability points λ_1 and λ_2 on the H_0 hypothesis, yielding ACVs respectively M_1 and M_2, c_1 is the better, or more sensitive, test if $M_1 < M_2$.

The line we take for the most part in this paper is perhaps more simple: if H_1 obtains (i.e. given $\mu > 0$) we equate c_1 found to its H_0 critical point, finding critical sample size n_1 therefrom. Similarly for c_2, finding sample size n_2. If $n_1 < n_2$, c_1 is more efficient.

Anything can happen with one particular experiment, e.g. we can find that k identifies relationship while r does not, while, in the long run, by the procedure indicated in the last paragraph, r may be found to be the more efficient. Consequently, in what follows, attention is directed, not to the single case, but to the overall showing, involving population values of the parameters, and not sampling values. This overall concept may be found difficult to understand: if so the reader may regard sample size as large (though we do not always in what follows). Anyway, conclusions will be found to be so decisive as to render meticulousness unnecessary.

[1] R. C. Geary, 1966, "The average critical value method for adjudging relative efficiency of statistical tests . . . ," *Biometrika* 53, 1 and 2, pp. 109–119.

[2] J. Neyman and E. S. Pearson, 1933, "On the Problem of Statistical Hypotheses", *Phil. Trans.* A, 231, pp. 289–337.

Let the critical null-hypothesis probability points (NHPP, e.g. .95, .99) be ρ_c and κ_c for estimates r and k computed from random samples of n. If there is, in fact, no relationship (hypothesis H_0) with tests r and k, decision will be right in the stated probability number of cases in the long run; both tests are equally efficient. But suppose that relationship is present (hypothesis H_1) in the population and that the variable pairs (X', Y') are normally distributed with $\rho = \rho_c (> 0)$. With sample size n not too small, decision as to presence of relationship in the population would be right in about half the cases in the long run, i.e. the power of the test is about 0.5.

This procedure (and conception generally) is repeated independently with the test k from the population (X, Y) (all 1 or 0), with population proportion of similars κ_c, same critical probability, same sample size, culminating in same power, approximately 0.5. Let κ' be population proportion corresponding to ρ_c, i.e. from (3)—

$$\kappa' = \frac{1}{2} + \sin^{-1} \rho_c. \tag{4}$$

The ACVs are here κ_c and κ', the latter "representing" r in being functionally related, by (4) to its ρ_c. If $\kappa' < \kappa_c$, r is a more efficient test of relationship than k because, using r, we would be right in our decision as to population relationship, whether absent or present, in more cases than if we used k.

Example 1

Let $n = 30$. actual .95 H_0 critical probability point $\rho_c = .361$. Hence—

$$\kappa' = \frac{1}{2} + \frac{1}{\pi} \sin^{-1} .361 = .618$$

As to the k-test in the H_0 case, with point-binomial symmetrical and n as large as 30, we may safely use normal theory (see note [3]), so that k will be regarded as normally distributed, mean 1/2 variance $\sigma \overset{2}{=} 1/(4 \times 30)$ (the familiar pq/n in this case).

[3] Approximate null-hypothesis ($\rho = 0$) var $r = 1/n$. To show that actual ·95 critical points are close to normal approximations take $n = 30$—

	r	k
Actual critical point	·361	·670*
Normal approximation	·358	·679

*Estimated by interpolation from point-binomial table, $p = 1/2$, $n = 30$.

Then—

$$\kappa_c = \frac{1}{2} + 1.96 \times .091287 = .679$$

Since $\kappa' < \kappa_c$, r is a more efficient test of relationship than k.

In the basic paper[1], as in the foregoing example, we postulated same sample size for the two tests being compared. From now on (to dramatise, so to speak, the difference in efficiency), we equalise the ACVs (in our particular case we set $\kappa' = \kappa_c$, having regard to relationship (4) between κ' and ρ_c) and compare the sample sizes required to bring about this equality. In the example, with $n = 30$ for test we find sample size n_k for k by setting $\kappa' = \kappa_c$, i.e.—

$$.618 = \frac{1}{2} + \frac{1.96}{2\sqrt{n_k}}$$

or $n_k = 69$, more than twice the size as if test r were used. In simple terms: using k instead of r would entail more than twice the cost of field work for a given level of sampling error tolerance.

Measurement of Superiority of Test

This means that ACV treatment can be used here in a more suitable way than in the original paper to measure precisely the superiority of r as a test of population $\rho \neq 0$. For what follows we calculate for different values of population parameter ρ the size of random sample which would be required by each test to decide that the right decision (i.e. that $\rho > 0$) would be made in about half the number of cases, on indefinitely large replication.

Since in the H_0 case we are dealing with symmetrical populations and not too small samples, it will be convenient to assume that normal theory applies for both tests, i.e. $N(0, 1/n)$ for r and $N(1/2, 1/4n)$ for k[3]. We require the H_0 normal critical .95 and .99 probability points, 1.96 and 2.5759. We also required, from (3),—

$$\delta = \kappa - \frac{1}{2} = \frac{1}{\pi} \sin^{-1} \rho_c$$

Let sample sizes be n_ρ and n_κ, then—

.95 probability

Test r: $\quad \dfrac{1.96}{\sqrt{n_\rho}} = \rho_c; \; n_\rho = \dfrac{(1.96)^2}{\rho_c^2} = \dfrac{3.8416}{\rho_c^2}$

Test k: $\quad \dfrac{1.96}{2\sqrt{n_\kappa}} = \delta; \; n_\kappa = \dfrac{(.98)^2}{\delta^2} = \dfrac{0.9604}{\delta^2}$

.99 probability

Test *r*: $$\frac{2.5759}{\sqrt{n_\rho}} = \rho_c;\ n_\rho = \frac{(2.5759)^2}{\rho_c^2} = \frac{6.6353}{\rho_c^2}$$

Test *k*: $$\frac{2.5759}{2\sqrt{n_\kappa}} = \delta;\ n_\kappa = \frac{(2.5759)^2}{4\delta^2} = \frac{1.6588}{\delta^2}$$

Hence, independent of probability level,—

(6) $$n_\kappa = \rho_c^2 n_\rho/4\delta^2 = \xi\, n_\rho;\ \xi = \rho^2/4\delta^2,$$

defining ξ.

Critical values for nine values of ρ are shown in the table, based on the foregoing formulation. By reference to $\rho = .1$, the table means that if population parameter $\rho = .1$ then a sample of 384 would be needed for test *r* to identify $\rho \neq 0$ correctly in half number of replications, but a sample of 948, $2\frac{1}{2}$ times as many, would be required using *k*, at null-hypothesis probability level .95.

TABLE: *Comparative Efficiency of Test Functions r and k for Different Population Values of* ρ

ρ	*Size of sample for probability*				$\xi = n_k/n_\rho$
	·95		*·99*		
	n_ρ	n_k	n_ρ	n_k	
·1	384	948	664	1,637	2·47
·2	96	234	166	404	2·44
·3	43	102	74	176	2·39
·4	24	56	41	97	2·33
·5	15	35	27	60	2·25
·6	11	23	18	40	2·14
·7	*(a)*	16	14	27	2·01
·8	*(a)*	11	10	19	1·84
·9	*(a)*	*(a)*	*(a)*	13	1·59

(a): fewer than 10.

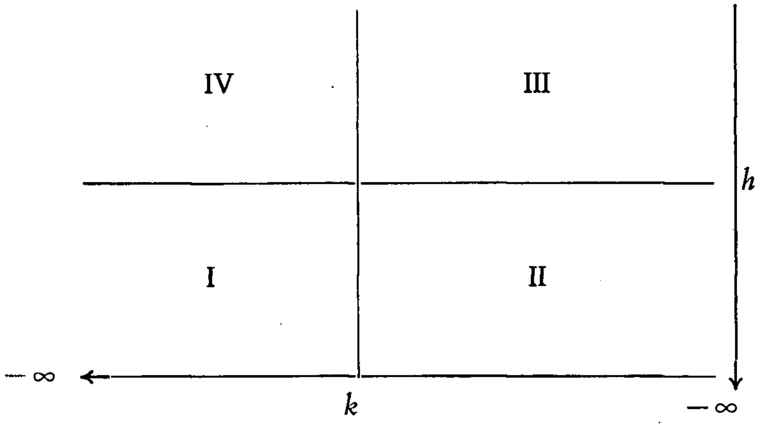
The last column (for ξ—see formula (6) above), independent of critical probability level, shows that, for e.g. population value = .1, about $2\frac{1}{2}$ times as large a sample (at $2\frac{1}{2}$ times field cost of survey) would be required to attain a given level of efficiency of decision if we were so foolish as to use *k* as a test instead of *r*. It is true that the relative efficiency improves as ρ increases, but slowly: if ρ were as large as .9999 the ratio ξ would still exceed unity, in fact 1.02; of course if ρ were 1 exactly the tests would be trivially equal in efficiency.

The 2× 2 Case

The simple situation considered above had the advantage that it could be worked in algebraic, i.e. general, terms; it could scarcely be regarded as of practical significance. The very familiar 2×2 case is different. The trouble is that the writer is unable to deal with it, from the present point of view, algebraically—a more ingenious colleague might like to take the problem up—and has to have recourse to constructed examples.

The bivariate normal distribution ((1) above) is again used, this time to construct a full 2×2 table, i.e. with 4 cell entries. Cell entry in quadrant 1—see diagram—will be, for h and k given,—

$$\text{(7)} \qquad \text{I} = \int_{-\infty}^{h} \int_{-\infty}^{k} f(x, y)\, dx\, dy$$



With—

$$\text{(8)} \qquad P(X) = \int_{-\infty}^{X} f(x)\, dx$$

the one variate normal integral, the other cell-entries are found as follows—

$$\text{(9)} \qquad \begin{aligned} \text{II} &= P(h) - I \\ \text{III} &= 1 - P(k) - II \\ \text{IV} &= P(k) - I \end{aligned}$$

Of course, the cell entries add to unity. The value of bivariate integral I was calculated from D. B. Owen's tables.[4]

[4] D. B. Owen, *Handbook of Statistical Tables*. Addison Wesley Publishing Company, Inc., 1962: pp. 184–204.

Example 2

$$\rho = .2,\ h = 1.2,\ k = 1.121816$$

The values of ρ and h were arbitrary. The value of k was taken to make reference to tables easier. The 2×2 table, ordered as in the diagram is—

.0901	.0249	.1150
.7789	.1061	.8850
.8690	.1310	1

Multiplying each entry by n, we regard the result as a random 2×2 sample of n. to be determined. This we do by equating χ^2 (with 1 d.f.) to its .95 and .99 probability points, as follows—

Prob.	NHCPP* 1 d.f.		χ^2	n
.95:	3.841	=	.008275n	464
.99:	6.635			802

*Null-hypothesis (H_0) critical probability point.

The picture we have then is of drawing an indefinitely large number of samples, always of size n, from an infinite population of which the unitary values are as shown in the foregoing 2×2 table. With n large enough the values found in the four cells for each sample will be near these population values (×n). About half the values of χ^2 found will be greater than the value shown (right decision following), about half less. We select this value (the NHCPP) as the average or typical value and equate it to the H_0 critical points to find sample size.

The r test case is simple. With $\rho = .2$ all sample values will be near .2 which, normalised, is equated to the NHCPP to find sample size.

Prob.	NHCPP (normal)		Deviation (.2) ÷stand. dev.	n
.95	1.96	=	$.2\sqrt{n}$	96
.99	2,5759			165

Example 3

Before comment, we give the results of another experiment of the same type, this time with population $\rho = .4$. Data—

$$\rho = .4,\ h = 0.7,\ k = 0.600781$$

The unitary 2×2 matrix is—

.1292	.1128	.2420
.5966	.1614	.7580
.7258	.2742	1

For χ^2—

Prob.	NHCPP (1 d.f.)		χ^2	n
.95:	3.841	=	$.058958n$	65
.99:	6.635			113

For r

Prob.	NHCPP (normal)		Deviation (.4) ÷stand. dev.	n
.95:	1.96	=	$.4\sqrt{n}$	24
.99:	2.5759			42

The Problem

There is no point in multiplying these examples until the validity of the viewpoint is fully discussed. The examples suggest that χ^2 (essentially an ordinal concept) is very insensitive compared to the cardinal (r) approach: example 2 shows that five times as large a sample would be required using χ^2 as using r (460 compared to 96 for .95 probability) to enable one to decide with confidence that relationship probably existed. If not so emphatic in example 3 (65 compared with 24), the showing is unmistakable.

While it is true that in the large majority of applications of χ^2 to the thousands of 2×2 experiments, it is not possible to conceive of cardinal measurement, one is left with the impression that in many cases, after possibly onerous and costly experimentation, researchers failed to find relationship when it was really present.

The problem which, without further ado, the author poses to his colleagues is as follows: can one cardinalise *every* 2×2 problem, i.e. convert it into a correlation problem and base deduction thereon? The procedure would reverse that of examples 2 and 3. Given a sample, with sample size not small, margins would enable the calculation of h and k, regarded, as in χ^2 theory, as given. A table giving values of χ^2, assuming normal bivariate theory, might be constructed in required detail of h, k, ρ. For given sample one would calculate χ^2 and read from the table the corresponding value of r (h, k, χ^2).

There exist tables [5], [6] from which, perhaps with recourse to interpolation, it should be possible to derive r fairly accurately from the unitized 2×2 table; it would appear that a further step is required, namely to effect the transition to χ^2, at least in the interim period prior to acceptance of the r approach for significance testing in every 2×2 case.

The Economic and Social Research Institute,
Dublin.

[5] K. Pearson (editor), *Tables for Statisticians and Biometricians, II* (1931), Biometric Laboratory, University College, London.

[6] U.S. Department of Commerce, "Tables of the Bivariate Normal Distribution Function and Related Functions", *National Bureau of Standards, Applied Mathematics*, 1959, Series 50.