THE ECONOMIC
AND SOCIAL
RESEARCH INSTITUTE
MEMORANDUM SERIES
NO. 118

RANDOMISATION AND THE VON NEUMANN FUNCTION: A VARIANCE FORMULA AND A PROBLEM

R.C. Geary.

Confidential: Not to be quoted
until the permission of the Author
and the Institute is obtained.

Randomisation and the von Neumann Function: A Variance Formula and a Problem

R.C. Geary.

In a paper of many years ago (Geary 1952) what was termed the contiguity ratio was introduced, to determine whether, in probability, a statistical map has a pattern or whether the mapped statistics are distributed at random. This ratio is really a two-dimensional version of the von Neumann (1941) statistic, more familiar as that tabled for null-hypothesis normal OLS residuals by J. Durbin and G.S. Watson. Geary was also concerned with the OLS residual problem. He approached it in two ways, by randomisation and by classical OLS regression theory, his instruments being means and variances of the contiguity ratio.

A difficulty with randomisation treatment was expressed as follows:-

> "The problem is to determine if there is a contiguity effect, i.e. if c [the contiguity ratio] has a significantly low value after the elimination of q independent variables by the least square method. As far as randomization is concerned, it would appear that the test developed in this section can be applied formally, the z being the remainders after the contributions of the independent variables have been removed. To a certain extent the writer shares the misgivings of some other students about the validity of the randomization approach in its application to regression remainders. As each successive independent variable is removed, should not the degree of freedom be diminished? It does not seem so. What happens is that the variance (or range) of the remainders diminish as the effect of each independent variable is allowed for, the test becoming indeterminate when the number of independent variables (originally with mean zero) is one less than the number of observations n, i.e. when all the remainders are zero. Accordingly the formal application of the randomization procedure, without diminution of the number of degree of freedom, does not result in obvious inconsistency: we can conceive of cases where c will be significantly low even after removal of the effect of $(n - 2)$ independent variables. Since doubts remain, however, the writer considered it desirable to examine the problem from the classical sampling aspect. In any case it will be interesting to compare the results of the two approaches. In the practical aspect the randomization method has the advantage that it can be applied without the assumption of universal normality in the n observations, regarded as a random sample."

As far as the writer is aware the degrees of freedom problem has never been discussed in this application: the controversy in another context between K. Pearson and R.A. Fisher is part of statistical history. One of the objects of the present communication is to invite statisticians to discuss the problem.

The contiguity ratio context is too esoteric for a suitable discussion. The problem arises in the much simpler single dimension of the ratio. But the writer is unaware of any randomisation treatment of the von Neumann statistic, so he ventures to give one here without any claim to originality. One result is remarkable, as will be seen.

Randomisation

One is given a sample of n measures of any kind (they may be raw values, OLS residuals etc), $x_1, x_2 \ldots, x_n$, ordered in a particular way. From a given function (e.g. the von Newmann ratio) one wants to make inferences about the character of the sample (is it probably non-normal, autoregressed etc ?). One considers the n! permutations of the sample values for each of which the test function has a value. These n! values are regarded as forming a frequency distribution. If the single value of the function found for the given ordered sample is near the ends of the frequency distribution (i.e. beyond the .05, .01 etc limits) one rejects the hypothesis, exactly as in ordinary theory. A feature of the test is that no assumption is made about the frequency distribution from which the sample of n is drawn. In theory one could calculate the moments of the function - or at least the first four moments - and so estimate the frequency distribution using e.g. the Pearson curve system. Here we deal only with the first two moments, the mean and the variance which suffice for most practical purposes.

The test function cannot usefully be symmetrical in $(x_1, x_2, \ldots, x_n)$ because then all the $n!$ values would be the same. The essence of the von Neumann ratio d is that it is not symmetrical (for $n > 2$) as it assumes that the sample elements are arrayed in a particular way. In fact, assuming, without loss of generality, that -

$$(1) \qquad \sum_{i=1}^{n} x_i = 0$$

as will always be the case with OLS residuals, d is given by -

$$(2) \quad d = \sum_{i=2}^{n} (x_i - x_{i-1})^2 \Big/ \sum_{i=1}^{n} x_i^2 = N/D$$

The numerator N is assymetrical, the denominator D symmetrical, i.e. D has the same value in all permutations. We need concern ourselves only with the numerator N. It is the fact of constant D that makes the calculation of moments of d exactly calculable. This is also the classical case when the sample is a normal one because then d is a homogeneous function of degree zero, with $nr^2 = \quad x_i^2$ in the denominator. The fact that when the sample is normal r is independent of d (Geary, 1933) makes the exact calculation of the moments of d, and hence the estimation of the frequency of d (as by Durbin-Watson) possible.

If f is any polynomial function of $(x_1, x_2, \ldots, x_n)$ ordered in a particular way the randomisation mean $M(f)$ of f is the sum of f for all the permutations divided by n! To find the mean of d^2 given by (2) or, in effect, N^2 we have to deal with terms in x_i^4, $x_i^3 x_{i'}$, $x_i^2 x_{i'} x_{i''}$ and $x_i x_{i'} x_{i''} x_{i'''}$, all subscripts different. On taking means we may disregard subscripts and insert mean values of these terms, having regard only to exponents. These mean values may be written (in a notation which is obvious) (4), (31), (22), (211), (1111). Note that (31) = (13) etc.

Square (1) and take means. There are n of type x_1^2 and $n\ (n - 1)/2$ of type $x_i x_{i'}$, $i' \neq i$. Hence -

(3) $$n\ (2) + 2\ \frac{n\ (n - 1)}{2}\ (11) = 0$$

or -

(4) $$(11) = -\ (2)/(n-1)$$

As in (4), we can express all terms - in two or more variables in single variable expressions. As an example of the method of derivation, we have

(5) $$\sum x_i\ \sum x_i^3 = 0.$$

Multiplying out and taking means -

(6) $$n\ (4) + n\ (n - 1)\ (31) = 0.$$

or -

(7) $$(31) = -\ (4)\ /\ (n - 1)$$

The derivation of other randomisation means we need is a little more complicated. We shall be content to give the results -

(8) $$\begin{aligned} (22) &= \left[n\,(2)^2 - (4)\right] / (n - 1) \\ (211) &= \left[2\ (4) - n\ (2)^2\right] / (n - 1)\ (n - 2) \\ (1111) &= 3\left[n\ (2)^2 - 2\ (4)\right] / (n - 1)\ (n - 2)\ (n - 3) \end{aligned}$$

From (2), -

(9) $$D = n\ (2).$$

Expanding the numerator of (2) -

(10) $$N = (x_1^2 + x_n^2) + 2 \sum_{i=2}^{n-1} x_i^2 \ - 2 \sum_{i=2}^{n} x_i x_{i-1}.$$

Hence taking means -

(11) $$M(N) = 2\ (2) + 2\ (n - 2)\ (2) + 2\ (n - 1)\ (2)/\ (n - 1),$$

using (4). Hence $M(n) = 2n\ (2)$. Then -

(12) $$M\ (d) = M\,(N)/D = 2,$$

using (9).

The algebra of the calculation of $M(d^2)$ or, in effect, $M\ (N^2)$ is onerous but the result is simple. We regard N, given by the right side of (10) as three terms $(A + B + C)$ with square $(A^2 + 2AB + \dots + C^2)$ and aggregate the terms, having regard to coefficients and numbers of terms of each kind, x_i^4, $x_i^3\, x_{i'}$ etc, which, on taking means are replaced by (4), (31) etc. Then, gathering terms we find -

(13) $$M\ (N^2) = 2\ (2n - 3)\ (4) - 8\ (2n - 3)\ (31) + 2\ (2n^2 - 4n + 3)\ (22)$$
$$- 8\ (n - 2)\ (n - 3)\ (211) + 4\ (n - 2)\ (n - 3)\ (1111).$$

Using (7) and (8) and collecting terms -

(14) $$M\ (N^2) = 2n\ \left[(2n^2 - 3)\ (2)^2 - (4)\right]\ /(n - 1).$$

As $M\ (d)^2 = M\ (N^2)\ /D^2$ with $D = n\ (2)$ -

(15) $$\begin{aligned} \text{var}\ (d) &= M\ (d^2) - \left[M\ (d)\right]^2 \\ &= 2\ \left[(2n - 3) - b_2\right]\ /n\ (n - 1) \end{aligned}$$

where $b_2 = (4)\ /\ (2)^2$ the familiar kurticity statistic in normal theory in which in fact its population value β_2 is 3.

As a check on the quite elaborate, if elementary, algebra, consider the case of $n = 2$. There is then but a single value of d given by (2), for in this case d is symmetrical in (x_1, x_2). $(4) = (x_1^4 + x_2^4)/2 = x_1^4$ since $x_1 + x_2 = 0$ and $(2) = x_1^2$. Hence $b_2 = 1$. Substituting then $n = 2$ and $b_2 = 1$ in (15) we find var $(d) = 0$, as we should.

Of course (15) is $O\ (n^{-1})$, which means that, with increasing n, d tends in probability towards 2 (see (12)). What, as announced above, is remarkable is that the coefficient of b_2 is $O(n^{-2})$. This implies that the variance is nearly independent of the frequency distribution from which the random sample of n (arrayed in any order) is drawn. As an example take $n = 20$ - one would scarcely be interested in e.g. residual autocorrelation for fewer observations - and $b = 1$, and 6, a range probably covering most distributions. Values of standard deviation (= square root variance) of var (d) given by (15) are -

Value of b_2	s.d. of d
1	0.4353
6	0.4039

The difference is of no importance, having regard to the uses of the statistic d.

Values of s.d. of d from $n = 20$ to $n = 100$ by tens with $b_2 = 3$, it normal value, are -

n	Standard deviation of d
10	0.5577
20	0.4230
30	0.3523
40	0.3080
50	0.2770
60	0.2538
70	0.2356
80	0.2207
90	0.2084
100	0.1980

The problem announced earlier remains. In this randomisation proceedure, does one have to take degrees of freedom into account in the most important application, namely in the study of OLS residual autocorrelation, and, if so, how?

Of course more than the variance is needed for the derivation of null hypothesis critical probably levels. For this at least the third and fourth randomisation moments of d would be required. After experience with the second moment, the writer surmises that the derivation of higher moments would be a prodigious task though perhaps approximations, say terms to n^{-2}

might not be too difficult. There does not seem to be much point in this exercise unless and until the degrees of freedom problem is cleared up; though the problem recedes in importance as sample size increases. Anyway, as practical researchers know, the twice s.d. deviation from mean suffices for most purposes of significance decision if one is not too particular about the probability involved. If one is, can at least appeal to Bienaymé - Tchebycheff! That the randomisation variance is practically distrubution - free is a powerful argument in its favour.

20 May 1977 R.C. Geary

References

R.C. Geary (1952): The contiguity ratio and statistical mapping, The Incorporated Statistician, Vol. 5, No. 3.

J. von Neumann (1941): Distribution of the ratio of the mean square successive difference to the variance, The Annals of Mathematical Statistics, XII, No. 4