ON RANDOMNESS IN ORDERED SEQUENCES

By L. C. Young

Westinghouse Electric and Manufacturing Company

It is frequently desirable to examine an ordered sequence of measurements for the presence of non-random variability, concern over any particular type of variability being limited. Unless the sequence is one containing replicated observations, current methods of analysis often restrict an investigation to tests for specific forms of variability, such as particular orders of regression and periodicity. In order to simulate replication, arbitrary grouping of data is occasionally used and followed by some test of variance; this practice, however, is likely to add an element of bias to the investigation.

Under tbese conditions, it would be convenient to have the means of testing a series for the presence of general regression, before proceeding to test for that of a specific type. It is the purpose of this paper to present, as briefly as possible, a statistic designed for this preliminary type of examination, and to demonstrate its application.

If a given sequence of measurements be denoted by

$$X_1, X_2, \cdots, X_n$$

then the magnitude of

$$C = 1 - \frac{\sum_1^{n-1} (X_i - X_{i+1})^2}{2 \sum_1^n (X_i - \bar{X})^2},$$

will be dependent upon the arrangement of the n observations upon which it is based. C will have $n!$ possible values for a given sample, corresponding to the number of permutations of n items.

1. Moments of the distribution of C in terms of the moments of a finite sequence.

Writing C in terms $x_1, \cdots, x_n$, representing the deviations of $X_1, \cdots, X_n$ from their sample mean of n measurements,

$$C = 1 - \frac{\sum_1^{n-1} (x_i - x_{i+1})^2}{2 \sum_1^n x_i^2}$$

$$= \frac{x_1^2 + x_n^2 + 2 \sum_1^{n-1} x_i x_{i+1}}{2 \sum_1^n x_i^2}.$$

In order to find the mean value of C for a given sample, it must be summed over all values obtained from the $n!$ permutations of the measurements.

Dealing with the numerator alone of the expression given above:

$$\sum\nolimits_p \left[x_1^2 + x_n^2 + 2\sum_1^{n-1} x_i x_{i+1}\right] = \sum\nolimits_p x_1^2 + \sum\nolimits_p x_n^2 + 2\sum\nolimits_p \sum_1^{n-1} x_i x_{i+1},$$

where Σ_p denotes summation over the $n!$ permutations.

There are n values of x_i, and $n!$ arrangements. Each value x_i is x_1 in $(n-1)!$ of the arrangements: the same reasoning applies to x_n. The first two terms of the summation, therefore, will be

$$\sum\nolimits_p x_1^2 = \sum\nolimits_p x_n^2 = (n-1)!\sum_1^n x_i^2.$$

With regard to the third term, there are $2(n-1)$ of such cross-products for each arrangement. Since the summation is taken over $n!$ arrangements, $x_j x_k$ will be different than $x_k x_j$, and should be considered a separate term. Each crossproduct term, therefore, must occur $\frac{2(n!)(n-1)}{n(n-1)}$ times throughout the $n!$ arrangements, since there are $n(n-1)$ possible cross-products among n different items. The third term, then, will be

$$2\sum\nolimits_p\left(\sum_1^{n-1} x_i x_{i+1}\right) = 2(n-1)!\sum_1^n\sum_1^{n-1} x_j x_k = -2(n-1)!\sum_1^n x_i^2,$$

from which it may be seen that the mean value of C is zero for any sample.

The same method may be applied in order to find the second and higher moments of C. Squaring the numerator of the expression and expanding,

$$\sum\nolimits_p\left[x_1^2 + x_n^2 + 2\sum_1^{n-1} x_i x_{i+1}\right]^2$$
$$= \sum\nolimits_p\left[x_1^4 + x_n^4 + 2x_1^2x_n^2 + 4x_1^2\sum_1^{n-1} x_i x_{i+1} + 4x_n^2\sum_1^{n-1} x_i x_{i+1} + 4\left(\sum_1^{n-1} x_i x_{i+1}\right)^2\right].$$

Performing the summation Σ_p term by term we obtain

$$\frac{\sum\nolimits_p\left[x_1^2 + x_n^2 + 2\cdot\sum_1^{n-1} x_i x_{i+1}\right]^2}{n!} = \frac{2(2n-3)\left(\sum_1^n x_i^2\right)^2 - 2n\sum_1^n x_i^4}{n(n-1)},$$

whence the second moment of C for any sample is given by

$$M_2 = \frac{2n - 3 - m_4/m_2^2}{2n(n-1)},$$

where m_2 and m_4 are the second and fourth moments, respectively, of the n observations about their mean.

In like manner, the third and fourth moments of the distribution of C for a given sample of n observations are found to be

$$M_3 = \frac{-6 + 4(n-3)\dfrac{m_3^2}{m_2^3} + 9\dfrac{m_4}{m_2^2} - 3\dfrac{m_6}{m_2^3}}{4n(n-1)(n-2)},$$

$$M_4 = \frac{1}{8n^3(n-1)(n-2)(n-3)}\left[24n^2(n-3)^2 - 48n(4n-9)\frac{m_3^2}{m_2^3}\right.$$

$$- 24n(3n^2 - 17n + 27)\frac{m_4}{m_2^2} + (8n^3 - 45n^2 - 23n + 210)\frac{m_4^2}{m_2^4}$$

$$+ 16(2n^2 + 5n - 21)\frac{m_5 m_3}{m_2^4} + 4(17n^2 - 37n + 42)\frac{m_6}{m_2^3}$$

$$\left. - (7n^2 + 13n - 6)\frac{m_8}{m_2^4}\right].$$

2. Distribution of C for samples drawn from a normal universe. The first four moments of the distribution of C for samples drawn from a given population may be derived from the above formulae by substituting the mean values of $\frac{m_3^2}{m_2^3}$, $\frac{m_4}{m_2^2}$, etc. of samples from such a population. For normal samples containing n observations, for example, the following mean values apply, as obtained by the method presented by R. A. Fisher [1, 2]:

$$\frac{m_3^2}{m_2^3} = \frac{6(n-2)}{(n+1)(n+3)},$$

$$\frac{m_4}{m_2^2} = \frac{3(n-1)}{(n+1)},$$

$$\frac{m_4^2}{m_2^4} = \frac{3(3n^3 + 23n^2 - 63n + 45)}{(n+1)(n+3)(n+5)},$$

$$\frac{m_5 m_3}{m_2^4} = \frac{60(n-1)(n-2)}{(n+1)(n+3)(n+5)},$$

$$\frac{m_6}{m_2^3} = \frac{15(n-1)^2}{(n+1)(n+3)},$$

$$\frac{m_8}{m_2^4} = \frac{105(n-1)^3}{(n+1)(n+3)(n+5)}.$$

Replacement of the sample moment ratios by the mean values of those ratios for normal samples yields the following moments of C:

$$M_1 = 0, \qquad M_2 = \frac{n-2}{(n-1)(n+1)}, \qquad M_3 = 0,$$

$$M_4 = \frac{3(n^2 + 2n - 12)}{(n-1)(n+1)(n+3)(n+5)}.$$

Compatible results for the case of normal samples have been obtained by Williams [3], using another method.

From the above results, the value of

$$\beta_2 = \frac{3(n^2 + 2n - 12)(n - 1)(n + 1)}{(n - 2)^2(n + 3)(n + 5)},$$

is seen to approach normality as the sample size is increased.

Inasmuch as the distribution of C for normal samples is limited in both directions and is symmetrical, it is apparent that the Pearson Type II distribution may be considered representative. Fitting this curve to the moments given above, the equation of the frequency distribution is given by

$$y = y_0\left(1 - \frac{C^2}{a^2}\right)^m,$$

where

$$m = \frac{(n^4 - n^3 - 13n^2 + 37n - 60)}{2(n^3 - 13n + 24)},$$

$$a^2 = \frac{(n^2 + 2n - 12)(n - 2)}{(n^3 - 13n + 24)},$$

$$y_0 = \frac{\Gamma(2m + 2)}{a \cdot 2^{2m+1}[\Gamma(m + 1)]^2}.$$

The values of β_2 for the distribution, for various values of n, are as follows:

Sample size, n	β_2
5	2.300
10	2.570
15	2.684
20	2.750
25	2.793
50	2.833

Due to the effect of even moments higher than the fourth, the approximation afforded by the Type II curve is not reliable for samples containing less than about eight observations. As the sample size decreases below this limit, the extremes of the C distribution deviate increasingly from the extremes ($\pm a$) of the fitted curve: with such a platykurtic distribution, therefore, the effect upon the lower significance levels vitiates the approximation.

Although either β_2 or the theoretical limits of the distribution of C could have been employed as a parameter of the fitted curve, it was considered expedient to use the former. In any case, of course, the advantage to be gained would be in connection only with samples containing few observations (less than eight). The evidence afforded by empirical sampling indicates that use of the limits as a parameter might render the approximation less valid.

In order to facilitate use of the approximate distribution for samples of eight or more observations, the values of C associated with two probability levels are tabulated below in Table I. The ratio of each value of C to its standard error is also shown, to demonstrate the approach to normality. The significance levels recorded exclude 10% and 2% of the area under the curve, respectively. In most practical applications, these will be the 5% and 1% levels, respectively, since only positive values of C exceeding the tabulated value will ordinarily be considered significant. The tabulations were prepared from tables of the function $I_x(p, q)$ [5], where $q = .5$ and $p = m + 1$, with the transformation $x = 1 - \frac{C^2}{a^2}$.

TABLE I

Significance levels of the absolute value of C

Sample size, n	$P = .10$	$C_{.10}/\sigma_c$	$P = .02$	$C_{.02}/\sigma_c$
8	.5088	1.6486	.6686	2.1664
9	.4878	1.6492	.6456	2.1826
10	.4689	1.6494	.6242	2.1958
11	.4517	1.6495	.6044	2.2068
12	.4362	1.6495	.5860	2.2161
13	.4221	1.6495	.5691	2.2241
14	.4092	1.6494	.5534	2.2310
15	.3973	1.6493	.5389	2.2369
16	.3864	1.6492	.5254	2.2423
17	.3764	1.6492	.5128	2.2470
18	.3670	1.6491	.5011	2.2513
19	.3583	1.6489	.4900	2.2550
20	.3502	1.6488	.4797	2.2585
21	.3426	1.6488	.4700	2.2616
22	.3355	1.6486	.4609	2.2647
23	.3288	1.6485	.4521	2.2676
24	.3224	1.6484	.4440	2.2700
25	.3165	1.6484	.4361	2.2717
Normal ($n = \infty$)		1.6447		2.3262

The distribution of C for normal samples containing 20 or more observations is sufficiently normal, for most practical cases and for the more common significance levels, to permit use of a table of areas under the normal curve, in conjunction with the standard error $\sigma_c = \sqrt{\frac{n - 2}{(n - 1)(n + 1)}}$. The 5% significance levels shown in Table I result, at worst, in a one per cent error of probability estimate, if the normal approximation is used in their place: that is, if 1.6447 times the standard error is used instead of the tabulated significance level, the probability will be .0505 at most, for the values of n which are tabulated.

3. General discussion on the application of C. It may be wondered why the statistic C has been used, rather than the more easily computed statistic

$$C' = \frac{\sum_1^{n-1} (X_i - X_{i+1})^2}{\sum_1^n x_i^2}.$$

As far as a significance test is concerned, it clearly does not matter which is used, since C and C' are linearly related. However, C may be regarded as symmetrically distributed about 0 in samples from a normal population to within at least four moments. Excessive departure of C from 0 may be taken as indicative of the presence of non-randomness in the series, the actual significance test being based, of course, on the probability of obtaining a departure larger than a given observed one, under the assumption of a random series. Positive values of C, in general, correspond to positive correlation while negative values correspond to negative correlation between successive observations.

There are various ways of detecting non-randomness in a series of observations, such as regression methods, analysis of variance, etc. The use of regression methods implies that we must know in general the type of regression function to be tried. C is a very flexible statistic, on the other hand, for testing the null hypothesis that a series is random, no matter what the alternative hypothesis is. A thorough study of C as a statistic for testing the hypothesis of randomness in an ordered series should include a study of the power function of C for hypotheses specifying various types of non-randomness. However, we shall simply appeal to intuition in proposing the statistic C, and forego power function considerations in this note. In practice, the advantage of using C increases with the length of a series: lack of randomness in a single sequence of ten or less observations may ordinarily be detected by regression methods, in fitting a low order polynomial. In a longer sequence of measurements, on the other hand, the presence of complicated regression or of periodicity is often sufficiently obscured by variation to elude detection by any other than a flexible method.

The statistic could be used to advantage in the field of applied statistics, in the investigation not only of variate series but of attribute series as well. For the latter purpose, an effort to tabulate the relationship between the level of significance and the percentage of either attribute would facilitate statistical investigation of random arrangement. A direct application could thus be made to binomially distributed attributes by a scalar assignment (0, 1) to the dichotomy, followed by a procedure similar to that presented above. Similarly, the randomness of vectorial observations could be examined from the viewpoint of arrangement. The common method of treating such problems,—the "random walk method,"—has occasionally been found inadequate in dealing with specific forms of non-random order; this is especially true when the allocable cause of variation has a multi-directional effect.

Needless to say, each of the fields of application considered so briefly above would require development before a routine, efficient method of investigating ordered arrangement could be established. Although probability level tables have been provided in this paper for C as applied to normal samples, it is quite evident that tables for samples from other parent distributions would be needed for some of the applications mentioned above.

4. An illustration of the use of C. Although one example has already been presented elsewhere [4] in which the distribution developed in Section 2 has been employed, a typical application of the statistic to an example in the field of quality control will be given here in order to illustrate the mechanics of solution. The data presented in Table II represent the percentages of defective product turned out daily, over a period of twenty-four days, by a single workman. The total output each day closely approximates five hundred parts: this fact is brought out to explain the calculation of χ^2 for the observed series of percentages, —it has no bearing upon the use of C.

TABLE II

Percentage of product rejected

Day	%, X	X^2	d^2
1	7.4	54.76	
2	8.8	77.44	1.96
3	11.4	129.96	6.76
4	10.3	106.09	1.21
5	11.9	141.61	2.56
6	12.2	148.84	.09
7	10.0	100.00	4.84
8	8.4	70.56	2.56
9	9.4	88.36	1.00
10	10.9	118.81	2.25
11	9.9	98.01	1.00
12	11.8	139.24	3.61
13	10.0	100.00	3.24
14	8.9	79.21	1.21
15	9.7	94.09	.64
16	9.3	86.49	.16
17	12.0	144.00	7.29
18	12.3	151.29	.09
19	10.3	106.09	4.00
20	8.6	73.96	2.89
21	10.4	108.16	3.24
22	11.1	123.21	.49
23	9.4	88.38	2.89
24	8.2	67.24	1.44
Totals	242.6	2495.82	55.42
		$n\bar{X}^2$ 2452.28	
		$\Sigma x^2 =$ 43.54	

$C = .3636$ (significant) $\chi^2 = 21.518$ (23 degrees of freedom) (not significant).

The value of C derived from the data lies between the two significance levels tabulated in Table I; there is reason to believe that the data are ordered, or nonrandom. Computation of χ^2, however, has been carried out with the hypothesis that all product was made under the same conditions (i.e. with a percentage defective equal to 10.108%, the mean of the group). The value so obtained is associated with a probability of about $P = .50$: the hypothesis is not disproved by this test. In short, the variability of the twenty-four observations could be considered random if it were not for the order of their arrangement.

REFERENCES

[1] R. A. Fisher, "Moments and product moments of sampling distributions," *Lond. Math. Soc. Proc.* (series 2) 30 (1929), pp. 199–238.

[2] R. A. Fisher, "The moments of the distribution for normal samples of measures of departure from normality," *Roy. Soc. Proc.*, A 130 (1930), pp. 16–28.

[3] J. D. Williams, "Moments of the ratio of the mean square successive difference in samples from a normal universe," *Annals of Math. Stat.*, Vol. 12 (1941), pp. 239-241.

[4] L. C. Young, "A critical appraisal of statistical methods in industrial management," presented at the annual meeting, American Society of Mechanical Engineers (1940).

[5] K. Pearson (Editor), *Tables of the Incomplete Beta-Function*, Biometrika Office, London 1924.