Reproduced from the Proceedings of the Royal Society, A, 125: 54-59 (1929), with permission of the Royal Society

75

TESTS OF SIGNIFICANCE IN HARMONIC ANALYSIS

Author's Note (CMS 16.53a) *

* Reproduced from Contributions to Mathematical Statistics (1950) by permission of John Wiley & Sons Inc.

A certain amount of disparity in usage and opinion among meteorologists wishing to test the significance of supposed periodicities had arisen in the absence of any treatment of this subject by the principles applicable to small samples.

The reader will notice that the periods of the components discussed in this paper are not integral multiples of the unit interval of observation, but integral submultiples of the entire period of observation. Using these it is shown that an exact solution can take account of the two circumstances which had given trouble (i) that the supposed periodicity to be tested is usually selected as having the largest apparent amplitude, and (ii) that the variability of the series must be estimated from the data, and needs to be eliminated from the test of significance.

The treatment of a frequency distribution consisting of a series of polynomial arcs with the highest possible contact at the points of discontinuity is of some mathematical interest.

I have also used this example to illustrate the properties of estimates involving non-linear functions of the frequencies **(Paper 163).**

The exact tests arrived at were embodied in a short table, designed to meet immediate and practical needs. A fuller table is published with this edition, giving for all values of n the number of submultiple periods available, from 5 to 50, both the 5 per cent and the 1 per cent values. The test is now so direct and easy that the urge to ascribe significance to casual fluctuations in time series, which all who deal with such series must have felt, should be capable of rational control.

Proceedings of the Royal Society of London, A, 125: 54-59, (1929).

Tests of Significance in Harmonic Analysis.

By R. A. Fisher, F.R.S.

(Received June 5, 1929.)

1. *Schuster's Test.*

If a series $u_1, u_2, \ldots, u_{2n+1}$ constitute a random sample from a normally distributed population, then any linear function

$$A = \overset{2n+1}{\underset{1}{S}} (a_r u_r)$$

will also be normally distributed; moreover its mean will be zero if $S(a_r) = 0$, and its variance will be equal to that of the original population if

$$S(a_r^2) = 1.$$

Any other linear function

$$B = \overset{2n+1}{\underset{1}{S}} (b_r u_r)$$

will be distributed independently of the first if

$$S(a_r b_r) = 0,$$

and in this case the sum of the squares,

$$x = A^2 + B^2,$$

will be distributed so that the chance of exceeding any particular value of x is

$$e^{-\frac{x}{c}},$$

where c is the mean value of x, equal to twice the variance of the population sampled.

This proposition, which gives the χ^2 distribution for the particular case $n = 2$, is the basis of Schuster's test of the significance of any particular term in the harmonic analysis of a series. For the coefficients

$$a_r = \sqrt{\frac{2}{2n+1}} \cos \frac{2\pi pr}{2n+1},$$

$$b_r = \sqrt{\frac{2}{2n+1}} \sin \frac{2\pi pr}{2n+1},$$

fulfil the necessary conditions for all integral values of p. Values of p from 1 to n give independently distributed values of x and, if the variance of the population were known *a priori*, the test would be rigorous for any one of these chosen in advance.

2. *Allowance for Selection of the Largest Term.*

The practice of picking out the larger values of x, not in advance, but by reason of their exceptional magnitude, requires, as Sir Gilbert Walker has shown, an important modification of the test of significance. For, if we wish to test the significance of the largest observed value of x, we must compare the value observed with the sampling distribution of the largest of n independent values, and not with that of any one value chosen in advance. If P stand for the probability which we adopt, as sufficiently small to be used as a criterion of significance, the corresponding value of x will be given by

$$e^{-\frac{x}{c}} = \mathrm{P},$$

for any particular term, but if x is chosen to be the largest of n independent values, it is necessary that the probability should be $1 - \mathrm{P}$ that all the n values shall be less than x, so that

$$\left(1 - e^{-\frac{x}{c}}\right)^n = 1 - \mathrm{P}$$

is the equation which determines the least value of x to be judged significant. This is the criterion derived by Walker.

3. *Allowance for the Sampling Error of the Estimated Variance.*

In the practical application of this criterion, when c is not known *a priori*, it is necessary to substitute for c an estimate of it derived from the data, and, for an exact test, to take into account the sampling error of this estimate. The estimate of c will necessarily be based on the variance observed in the original sample, or, what comes to the same thing, on the average value of x for the n possible periods; and, whether we take, as our actual estimate, the average of all the n values, or the average of the $(n - 1)$ values other than that to be tested, all that is required for an exact solution, in either case, is the frequency distribution of the largest of n values of x, expressed as a fraction of the total of the sample of n of which it is the largest member.

If $x_1, \ldots, x_n$ are the co-ordinates of a point in Euclidian space of n dimensions, the simultaneous distribution of the n values will be represented by a density function

$$e^{-\frac{1}{c}(x_1 + x_2 + \ldots + x_n)}$$

which is constant over plane finite regions of $n-1$ dimensions, bounded by the n-surfaces

$$x_r = 0$$

in the form of a generalised tetrahedron. In every such region, the distribution of the ratio of the largest co-ordinate to the sum of all co-ordinates will be the same, and, since the density is constant over each such region, the distribution is to be found merely from the elements of generalised volume, into which the region is divided for fixed values of the ratio. Any particular co-ordinate, *e.g.*, x_1, will be the greatest in one nth of the whole region, this fraction being bounded on the one hand by the loci, at which it ceases to be greatest,

$$x_1 = x_2, \quad x_1 = x_3, \quad \dots, \quad x_1 = x_n,$$

and, on the other, by the boundaries,

$$x_2 = 0, \quad x_3 = 0, \quad \dots, \quad x_n = 0\,;$$

within this region it is required to find the distribution of the ratio

$$g = \frac{x_1}{x_1 + x_2 + \dots + x_n}.$$

4. *The Discontinuities of the Distribution.*

The distribution defined geometrically by the dissection of a generalised tetrahedron exhibits a number of discontinuities; the linear regions which constitute its boundary intersect $n-1$ at a time at the sets of points typified by

$$\begin{array}{l} x_1 = x_2 = x_3 = \dots\dots\dots = x_n \\ x_2 = 0, \quad x_1 = x_3 = \dots\dots = x_n \\ x_2 = x_3 = 0, \quad x_1 = x_4 = \dots = x_n \\ \dots\dots\dots\dots\dots\dots\dots \\ x_2 = x_3 = \dots\dots\dots\dots = x_n = 0, \end{array}$$

at which it is evident that the values of g are

$$\frac{1}{n}, \quad \frac{1}{n-1}, \quad \dots, \quad \tfrac{1}{2}, \quad 1,$$

representing in succession, the centre of the generalised tetrahedron, the centres of all its bounding faces, of successively lower dimensions, which meet in the point, $g = 1$, the middle points of the edges running from this point, and finally the limiting point, $g = 1$, itself. Hence g is distributed over the range from $1/n$ to 1; and for an exact test of significance we require to know the probability with which any particular value between these limits is exceeded.

5. *The Exact Distribution.*

A point about the distribution which greatly facilitates the solution, is that within the region between any two discontinuities the probability integral of the distribution is merely a polynomial in g of degree $n - 1$. For the boundaries of any region, $g = g_0$, change the magnitude of their elements continuously at rates determined by the magnitude of their boundaries, and so on down to the bounding edges, the lengths of which are linear functions of g; consequently the probability integral is in each region a polynomial of degree $n - 1$, but from region to region the $(n - 1)$th differential coefficient with respect to g changes discontinuously.

We may therefore represent the probability integral by the form

$$\mathrm{P} = \alpha_1 (1 - g)^{n-1} + \alpha_2 (1 - 2g)^{n-1} + \ldots + \alpha_n (1 - ng)^{n-1},$$

in which as many terms are to be taken as have positive quantities within the brackets. The last term is therefore included for no possible value of g, but is written above in order to utilise the condition that when $g < 1/n$ the probability integral shall be unity. This condition is sufficient to determine the n coefficients by equation of the coefficients of $g^0, g^1, \ldots, g^{n-1}$.

To determine their actual values let

$$f = -1 + \alpha_1 t + \alpha_2 t^2 + \ldots + \alpha_n t^n,$$

then the equations of the successive coefficients give

$$f = 0, \quad \left(t \frac{d}{dt}\right) f = 0, \quad \ldots, \quad \left(t \frac{d}{dt}\right)^{n-1} f = 0,$$

at the value $t = 1$. These are evidently equivalent to

$$f = 0, \quad \frac{d}{dt} f = 0, \quad \ldots, \quad \frac{d^{n-1}}{dt^{n-1}} f = 0,$$

for the same value, so that f, being of degree n, must be a numerical multiple of $(t - 1)^n$, or, in view of its first term,

$$f = -(1 - t)^n.$$

We have therefore the probability integral in the form

$$P = n(1-g)^{n-1} - \frac{n(n-1)}{2}(1-2g)^{n-1} + \dots + (-)^{k-1}\frac{n!}{k!\,(n-k)!}(1-kg)^{n-1},$$

where k is the greatest integer less than $1/g$.

6. *Summary and Table of* 5 *per cent. Values.*

A practical convenience of the form which has been obtained for the probability integral, is that for small values of P, such as are needed in tests of significance, the magnitude of the successive terms decreases very rapidly, so that even when, as at the 5 per cent. point for $n = 50$, as many as seven terms exist, very high precision is obtained from the first three terms only. Indeed the first term alone gives a very satisfactory approximate test of significance. The first term has, moreover, a simple meaning in relation to a related statistical problem. There are, in fact four related distributions each of which is the appropriate solution of one of four problems.

(I) The distribution of any one harmonic term obtained from a random sample of numbers drawn from a population of known variance. Schuster's solution of this is given by the distribution of the form

$$P = e^{-x}. \tag{1}$$

(II) The distribution of the largest of the n harmonic terms obtained from a similar sample ; for this we have Walker's solution

$$P = 1 - (1 - e^{-x})^{n}. \tag{2}$$

(III) We may ask what is the distribution of any one harmonic term as a fraction of the total (or mean) of the terms obtained from the same sample ; here there is no restriction that our term should be the largest, and all points within the generalised tetrahedron are available, so that

$$P = (1 - g)^{n-1} \tag{3}$$

where g is the chosen term expressed as a fraction of the whole.

(IV) Finally the probability that the largest of the n terms should exceed g is, so long as this probability is small, naturally not far from n times the value given by (3), and has been shown to be exactly

$$P = n(1-g)^{n-1} - \dots + (-)^{k-1}\frac{n!}{k!\,(n-k)!}(1-kg)^{n-1}, \tag{4}$$

where k is the largest integer less than $1/g$.

How good an approximation is obtained by using the first term only, is shown by the following table giving the 5 per cent. values of g for values of n from 5 to 50 in a parallel column with those obtained by ignoring all terms after the first.

n.	g (by exact formula).	g (by first term only).
5	0·68377	0·68377
10	0·44495	0·44495
15	0·33462	0·33463
20	0·27040	0·27046
25	0·22805	0·22813
30	0·19784	0·19794
35	0·17513	0·17525
40	0·15738	0·15752
45	0·14310	0·14324
50	0·13135	0·13149

This table can be used directly in testing significance; the 5 per cent. point is the lowest level of significance likely to be wanted, and for higher levels, such as the 1 per cent. point, the first term will provide an even closer approximation. The method of section 5 should be useful in many distribution problems involving points of discontinuity.

The value of g may in all cases be very easily obtained. If all the Fourier submultiples have been worked out, it is, as already defined,

$$\frac{x_1}{x_1+x_2+\dots+x_n}.$$

The denominator of this expression is, however, merely

$$\mathop{\mathrm{S}}_{r=1}^{2n+1}(u_r-\bar{u})^2.$$

In the case where the number of observations in the series is even, $(2n+2)$, we need still only consider the n complete harmonic terms, and can obtain their sum as

$$\mathop{\mathrm{S}}_{r=1}^{2n+2}(u_r-\bar{u})^2-\frac{(u_1-u_2+u_3-\dots-u_{2n+2})^2}{2n+2}.$$

Author's revised version of table on preceding page (CMS 16.59a).

Significant Values for g, the Ratio of the Sum of Squares for the Most Significant Period N. (The Total for n Periods Obtainable from a Sequence of $(2n + 1)$ or $(2n + 2)$ Successive Observations.)

n	5%	1%	n	5%	1%	n	5%	1%
5	.68377	.78874	20	.27040	.32971	35	.17513	.21338
6	.61615	.72179	21	.26060	.31783	36	.17124	.20860
7	.56115	.66440	22	.25155	.30683	37	.16754	.20405
8	.51569	.61517	23	.24315	.29661	38	.16400	.19970
9	.47749	.57271	24	.23534	.28709	39	.16062	.19554
10	.44495	.53584	25	.22805	.27819	40	.15738	.19156
11	.41688	.50357	26	.22123	.26986	41	.15429	.18776
12	.39240	.47510	27	.21483	.26205	42	.15132	.18411
13	.37085	.44982	28	.20883	.25470	43	.14847	.18060
14	.35172	.42722	29	.20317	.24778	44	.14573	.17724
15	.33461	.40689	30	.19784	.24124	45	.14310	.17401
16	.31922	.38851	31	.19280	.23506	46	.14057	.17089
17	.30529	.37180	32	.18803	.22921	47	.13814	.16789
18	.29262	.35655	33	.18351	.22366	48	.13579	.16501
19	.28104	.34257	34	.17921	.21839	49	.13353	.16222
20	.27040	.32971	35	.17513	.21338	50	.13135	.15954

Table of g; **for testing the significance of the leading periodic component of a series of $2n + 1$ or $2n + 2$ consecutive values. Each of n components contributes a certain fraction to the total sum of squares, and g is taken to be the largest of these fractions. If this exceeds the corresponding tabulated value significant evidence of periodicity is indicated.**