# AN APPLICATION OF ANALYSIS SITUS TO STATISTICS*

BY HAROLD HOTELLING

1. *Introduction.* The theoretical distribution curve of correlation coefficients obtained by sampling under any particular conditions may or may not extend to the limits $r=\pm 1$. Whether it does extend to either of these limits, and if so its order of contact with the $r$-axis, are determined by features of the problem which will be shown to be essentially topological. These properties of the curve are independent of any influence which some members of the sample may exert upon others, provided this does not amount to complete determination. They are, moreover, to some extent independent of the nature and existence of any correlation between the variates in the population from which samples are drawn, and indeed of the distribution in this population of the variates. Finally, they are independent of heterogeneity in the population. First it will be well to notice the relation of these ideas to *time series.*

2. *Time Series.* It is well known that the correlation coefficient and other statistical measures do not have their usual significance when computed from observations ordered in time. On account of the lack of independence of successive observations, the known sampling distributions and probable error formulas are inapplicable, and no adequate substitutes have been discovered. Economic and social statisticians and meteorologists thus labor under a serious handicap which is largely absent from biometric work.

A contribution toward the removal of this handicap has recently been made by G. Udny Yule in a presidential address to the Royal Statistical Society.† He gives the

* Presented to the Society, San Francisco Section, October 30, 1926.

† *Why do we sometimes get nonsense correlations between time series?* Journal of the Royal Statistical Society, vol. 89 (1926), pp. 1–69.

name *conjunct series* to sequences whose finite differences of any definite order may be considered independent, and inquires as to the distribution of correlation coefficients between conjunct series. Failing to discover a mathematical solution, he resorts to experiment, and gives a frequency distribution of 600 correlation coefficients between conjunct series obtained by cumulating in sets of ten the results of drawing numbers at random from a box. He fits the frequency distribution by an empirical curve, but with the surmise that this does not represent the true function. The correctness of this surmise now appears from the fact that the curve has vertical tangents at the ends of the range, whereas, if the conditions of the experiment be slightly idealized in order to yield an analytic curve (or any curve having a sufficient number of derivatives), this will be horizontal at the ends, and further, will have second-order contact with the axis. Yule's data as presented in his Figure 17 may indeed suggest contact with the axis.

A somewhat different class of questions arises in examining the conformity of a historical variable to a differential equation. Consider for example equations of the form

$$\frac{d^m x}{dt^m} = a + bx,$$

which occur in connection with population estimates, growth curves and studies of periodicity. Certain logical advantages, which are dwelt on elsewhere, attach to the following procedure. From a sequence of observed values of $x$ the values of the $m$th derivatives at the corresponding times are estimated by means of finite differences. The differential equation then becomes a regression equation: $a$ and $b$ may be determined by least squares, and the correlation coefficient used to test the goodness of fit. But even if the differential equation were without validity and the values of $x$ or of $d^m x/dt^m$ mere random numbers, the correlation coefficients thus obtained would not have the distribution nor probable error ordinarily attributed to $r$. For here one set of quantities is derived by manipulation of the other rather than by independent observation.

These and many similar outstanding problems in time series seem to be of great mathematical difficulty, and it is likely that investigators will often have to depend for assurance of the validity of their results upon rough ideas derived from experiments with cards and dice.

3. *Geometric Relations.* Letting the observations $x_1, x_2, \cdots, x_n$ be cartesian coordinates of a point $P$, the operation of replacing them by deviations from their mean is equivalent to projecting $P$ orthogonally upon the hyperplane

$$x_1 + x_2 + \cdots + x_n = 0.$$

Let the projected point be projected radially from the origin $O$ upon the hypersphere

$$x_1^2 + x_2^2 + \cdots + x_n^2 = 1.$$

This represents a mere change of units for the observations, which does not alter their correlation with any other sequence. If the observations are independent random drawings from a normally distributed aggregate, they are represented by a point $Q$ taken at random on the hypersphere in such a way that the element of probability is proportional to the element of $(n-2)$-dimensional volume. But since we are not requiring the drawings to be independent, random nor from a normally distributed aggregate, we assume only that the probability that the representing point will be found within a distance $\delta$ of a point $Z$ of the hypersphere, divided by $\delta^{n-2}$, approaches a continuous positive function of $Z$ as $\delta$ approaches zero.

If $Q$ and $S$ are two points on the hypersphere the coefficient of correlation between the corresponding sequences equals cos $QOS$, as is well known. If $S$ is determined from $Q$ in a definite manner, this method of determination defines a transformation $T$ of the hypersphere into itself or a part of itself, and we are interested in the amount of motion involved. The correlation is now a function of position; the hypersphere will be marked out with hypersurfaces of constant correlation, and it is the volume between such hypersurfaces multiplied by a continuous positive function, that determines our frequency curve of sample correlations. The case of independent selection of the points may be regarded as the special and extreme case in which the whole sphere is transformed into one point.

4. *Transformations of a Hypersphere.* A topological study of transformations of an $m$-dimensional sphere into itself has been made by J. W. Alexander,* who has proved. among other things, that a one-to-one continuous transformation of a sphere into itself always has at least one fixed point if the transformation preserves sense and the sphere is of even dimensionality, or if the transformation reverses sense and the dimensionality is odd.

In a transformation $T$ of a sphere corresponding to a transformation of a statistical series, a fixed point corresponds to a correlation of $+1$. To a correlation of $-1$ corresponds a fixed point of a new transformation consisting of $T$ followed by a transference $R$ of each point to its diametrically opposite point. This transference $R$ preserves sense of orientation if the sphere is of an odd number of dimensions, but reverses sense if the number of dimensions is even. Alexander's theorem quoted above may therefore be interpreted as follows in terms of the curve of distribution of the correlation coefficient $r$ between a series and its transformed series. Let $n$ be the number of observations; the hypersphere is then of $n-2$ dimensions. If $n$ is even and $T$ preserves sense of orientation, it has a fixed point. The distribution curve for $r$ therefore extends to $+1$. If $n$ is even and $T$ reverses sense, $TR$ has a fixed point, so that the distribution curve extends to $-1$. If $n$ is odd and $T$ reverses sense the curve extends to $+1$. Only if $n$ is odd and $T$ preserves sense is it possible for the distribution curve to stop short of both extremes, for in this case only neither

* Transactions of this Society, vol. 23 (1922), pp. 89–95.

$T$ nor $TR$ is compelled to have an invariant point. These statements may be summarized in the following table.

| | $n$ even | $n$ odd |
|---|---|---|
| $T$ reverses sense | For some samples $r = -1$ | For some samples $r = +1$ |
| $T$ preserves sense | For some samples $r = +1$ | The curve may or may not extend to $r = \pm 1$. |

5. *Correlation with Permuted Series.* As an illustration consider the correlation between series $x_1, x_2, \cdots, x_n$ and $x_1', x_2', \cdots, x_n'$ connected by the permutation.

$$T: \qquad x_1' = x_2,\ x_2' = x_3,\ \cdots,x_{n-1}' = x_n,\ x_n' = x_1.$$

This is closely related to what Yule (loc. cit.) has called the serial correlation of first order. If $n$ is odd the substitution represents in $n$-space a rotation, the determinant being $+1$, and therefore preserves sense. If $n$ is even, sense is reversed. The same statements hold for the hyperplane

$$x_1 + x_2 + \cdots + x_n = 0,$$

which is transformed into itself; for each of the regions into which this hyperplane divides the $n$-space goes into itself. They must therefore hold for the unit sphere in this hyperplane. The situation is described by the upper left and lower right compartments of the table above. If $n$ is even we have, in fact, the sequence

$$1, -1, 1, -1, \cdots, 1, -1$$

which has a correlation $-1$ with its transformed sequence.

If $n$ is odd, we shall show that for this transformation the frequency curve for $r$ does not extend to either of the extreme values $\pm 1$; there is a maximum value less than unity which the correlation cannot exceed numerically. If this were not true it would follow from continuity considerations that, for $T$ or for $TR$, a fixed point on the $(n-2)$-dimensional sphere could be found. Indeed the distance from a point to its transform is a function varying continuously over a finite region, and must therefore take a minimum value. Suppose, then, that a sequence of real numbers $x_1, \cdots, x_n$ can be found which is perfectly correlated with its transformed sequence. We can then find real numbers $\rho$ and $\mu$ such that

$$\rho x_i + \mu = x_i' = x_{i+1}, \qquad (i = 1,2, \cdots, n-1),$$
$$\rho x_n + \mu = x_n' = x_1.$$

These are homogeneous linear equations for determining $x_1, \cdots, x_n$ and $\mu$, and have a matrix which, for $n=5$, is

$$\left\| \begin{matrix} \rho & -1 & 0 & 0 & 0 & 1 \\ 0 & \rho & -1 & 0 & 0 & 1 \\ 0 & 0 & \rho & -1 & 0 & 1 \\ 0 & 0 & 0 & \rho & -1 & 1 \\ -1 & 0 & 0 & 0 & \rho & 1 \end{matrix} \right\| .$$

The determinant of the first $n$ columns equals $\rho^n - 1$, which, since $n$ is odd, vanishes only for $\rho = 1$. But if $\rho = 1$ the determinant of the last $n$ columns equals $n$, as is seen by adding to the last row all the others. Hence the rank of the matrix is $n$. Hence all solutions of the equations are proportional to one solution. But one solution is $\mu = 1 - \rho$, $x_1 = x_2 = \cdots = x_n = 1$, so that the $x$'s must all be equal for every solution. No such point lies on our hypersphere; indeed the definition of the correlation coefficient presupposes that the $x$'s are not all equal.

The permutation, related to the second serial correlation,

$$x_1' = x_3, \; x_2' = x_4, \cdots, x_{n-1}' = x_1, \; x_n' = x_2$$

supplies an example of a transformation which always preserves sense.

6. *Singular Transformations.* For continuous transformations of the sphere into a part of itself, the distribution curve for $r$ must extend to both extremes. For, in Alexander's terminology, the index is zero, and there must therefore be

an invariant point. In this manner it may be shown that (as is also evident otherwise) a sequence may have a correlation of either 1 or $-1$ with its differences of any order.

7. *Dimensionality determines Order of Contact.* A further property of the distribution of $r$ can be inferred from the dimensionality of the locus of invariant points. Let this locus $\pi$ be of $p$ dimensions. Let $u_1, u_2, \cdots, u_{n-2}$ be coordinates on the hypersphere such that $u_{p+1}, u_{p+2}, \cdots, u_{n-2}$ are zero at all points of $\pi$ and which on the coordinate lines vary as the distance from $\pi$, apart from infinitesimals of higher order. We shall assume enough regularity in $\pi$ and $T$ so that the equations of the transformation may be written

$$u_1' - u_1 = a_{1,p+1}u_{p+1} + a_{1,p+2}u_{p+2} + \cdots + a_{1,n-2}u_{n-2} + \eta_1,$$
$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$
$$u_{n-2}' - u_{n-2} = a_{n-2,p+1}u_{p+1} + a_{n-2,p+2}u_{p+2} + \cdots$$
$$+ a_{n-2,n-2}u_{n-2} + \eta_{n-2},$$

where $\eta_1, \cdots, \eta_{n-2}$ are functions which vanish to the second or higher order with $u_{p+1}, \cdots, u_{n-2}$ and where the coefficients $a_{ij}$ are either constants or functions of $u_1, \cdots, u_p$ and are not all identically zero. This condition on $T$ will be satisfied for example if in a neighborhood of every point of $\pi$ the transformation $T$ can be approximated arbitrarily closely by a linear transformation. The transformation $T$ is not restricted to be one-to-one and the matrix of the approximating linear transformation may be of any rank. In an extreme case every point might go into a point of $\pi$. The absence from the right-hand members of the equations of terms not involving $u_{p+1}, \cdots, u_{n-2}$ is due merely to the fact that all the expressions $u_i' - u_i$ must vanish on $\pi$.

In the right-hand members put $u_i = sv_i$ $(i = p+1, p+2, \cdots, n-2)$, where $s$ is the distance from $\pi$. Since not all the $a_{ij}$'s are zero, at least one of the equations will have a term of the first degree in $s$.

The distance $k$ from a point to its transform differs only by terms of higher order from the square root of a homogeneous positive definite quadratic form in $u_1'-u_1$, $u_2'-u_2$, $\cdots$, $u_{n-2}'-u_{n-2}$, so that we may write

$$k = as + bs^2 + \cdots ,$$

the coefficients $a$, $b$, $\cdots$ depending on the direction from $\pi$, and $a$ being positive everywhere except possibly in isolated directions, where it may vanish. By integrating over the manifold for which $k$ is constant we find, denoting the mean value of $s$ by $\rho$,

$$k = \alpha\rho + \beta\rho^2 + \cdots ,$$

where $\alpha$, $\beta$, $\cdots$ are constants and $\alpha > 0$. This step is justified by the fact that the manifold is closed and everywhere close to $\pi$ if $k$ is small, which follows from the positive definite character of the quadratic form in $u_1'-u_1$, $\cdots$, $u_{n-2}'-u_{n-2}$ giving the fixed number $k^2$.

Since $\alpha \neq 0$, we may invert the last series, obtaining

$$\rho = Ak + Bk^2 + \cdots .$$

Now $\pi$ is the uniform limit of a family of closed non-intersecting $(n-3)$-dimensional loci, on each of which the distance $k$ from a point to its transform is constant. For $\pi$, $k=0$. The $(n-3)$-dimensional volume of one of these loci divided by the $p$-dimensional volume of $\pi$ is a function of $\rho$ whose Maclaurin expansion begins with a constant multiple of $\rho^{n-p-3}$. When this volume is expressed in terms of $k$ its expansion will therefore begin with a term in $k^{n-p-3}$. The $(n-3)$-dimensional volume is proportional to the ordinate of the frequency curve of $k$. Since $k$, when small, differs only by infinitesimals of higher order from $(1-r^2)^{1/2}$ and therefore from $(1-r)^{1/2}$, it follows that the frequency for $r$, when expanded about $r=1$, will begin with a term in $(1-r)^{(n-p-3)/2}(dk/dr)$, that is, in $(1-r)^{(n-p-4)/2}$. In this way the dimensionality of the locus of invariant points, which is a topological feature, determines the order of contact of the frequency curve with the axis at the upper extreme. In like manner the dimensionality of the locus for $TR$ determines the order of contact at $r=-1$. If the locus consists of several parts, $p$ refers to the part having the greatest number of dimensions.

8. *Parameter Forms.* The transformation $T$ may be considered not as a single transformation but as depending on parameters $t_1, t_2, \cdots, t_k$ which vary continuously. A frequency distribution of probability $\phi(t_1, t_2, \cdots, t_k)$ is then to be assumed for the parameters. Suppose the dimensionality $p$ of the invariant locus is the same for all values of the parameters. The probability of a correlation between $r$ and $r+dr$ in a sample of $n$, divided by $dr$, will be given by an expansion in powers of $1-r$ beginning with a constant times $\phi(t_1, \cdots, t_k)$ $(1-r)^{(n-p-4)/2}$, provided the parameters have the particular values $t_1, \cdots, t_k$. Without this proviso, the probability is found by integration with respect to $t_1, \cdots, t_k$ over their entire field. Since $r$ is constant with respect to this integration, and since $\phi$ is everywhere positive, it appears that the order of contact of the frequency curve for $r$ with the axis is the same as before.

9. *Correlation between Variates in general.* This enables us to apply the result of §7 to the correlation between variates of which one is not fully determined by the other. Let the point on the $(n-2)$-dimensional sphere determined by the $n$ values in a sample of the first variate be $Q$, and let $\phi(t_1, t_2, \cdots, t_{n-2})dt_1\, dt_2 \cdots dt_{n-2}$ be the probability that $Q$ shall have coordinates differing from $t_1, \cdots, t_{n-2}$ respectively by less than $dt_1, \cdots, dt_{n-2}$. We now regard $T$ as a transformation of the entire sphere into the point $Q$, which thus constitutes the invariant locus. Hence $p=0$ and the expansion in powers of $1-r$ begins with $(1-r)^{(n-4)/2}$. Contact is therefore of order $(n-2)/2$. If $T$ be preceded by the transformation $R$ of every point to its diametrical opposite it appears that the order of contact at $r=-1$ is the same.

These results agree with the known distribution*

* R. A. Fisher, *On the influence of rainfall on the yield of wheat at Rothamsted*, Philosophical Transactions of the Royal Society, vol. 213B (1923), p. 92. A more general distribution is given by Fisher in Biometrika, vol. 10 (1915).

$$\frac{1}{\pi^{1/2}}\frac{\Gamma[(n-1)/2]}{\Gamma[(n-2)/2]}(1-r^2)^{(n-4)/2}$$

of correlations in samples of $n$ drawn from a normally distributed aggregate in which the variates are uncorrelated.

When, as in Yule's experiment mentioned in §2, $n=10$, contact will of course be of second order.

10. *Extension to Multiple Correlation.* To extend the results of the last section to multiple correlation, let the $q$ points representing the independent variates on the hypersphere $S_{n-2}$ which is the intersection of

$$S_{n-1}: \qquad x_1^2+x_2^2+\cdots+x_n^2=1,$$

and

$$V_{n-1}: \qquad x_1+x_2+\cdots+x_n=0,$$

determine with the origin a hyperplane $V_q$ within $V_{n-1}$. Let $S_{q-1}$ be the hypersphere of $q-1$ dimensions in which $V_q$ meets $S_{n-2}$. The multiple correlation coefficient $R$ will be unity if the point representing the dependent variable lies on $S_{q-1}$. Here $p=q-1$; the series at $R=1$ will therefore begin with a multiple of $(1-R)^{(n-q-3)/2}$.

The other end of the range of the multiple correlation coefficient is at $R=0$. On $S_{n-2}$ the locus of points for which $R=0$ is of $n-q-2$ dimensions. The $(n-3)$-dimensional volume at distance $R$ from this locus is proportional to $R^{(n-3)-(n-q-2)}=R^{q-1}$, apart from higher powers of $R$. As a special case we have Fisher's formula, given in a slightly different form in the 1923 paper cited,

$$\frac{\Gamma[(n-1)/2]}{\Gamma[(n-q-1)/2]\Gamma(q/2)}R^{q-1}(1-R^2)^{(n-q-3)/2}$$

for the distribution of $R$ in random samples from uncorrelated material, a formula which for $q=1$ reduces to that for $r$.

Stanford University