ON CERTAIN TOPICS IN THE MATHEMATICAL THEORY OF STATISTICS*

BY H. L. RIETZ

1. *Introduction.* The mathematical theory of statistics dates back to the first publication† relating to Bernoulli's theorem in 1713. The line of development started by Bernoulli was carried forward by DeMoivre,‡ Stirling,§ Maclaurin,|| and Euler¶ culminating in the formulation of the Bernoulli theorem by Laplace** in substantially the form in which it still holds a fundamental place in mathematical statistics.

The *Théorie Analytique des Probabilités* of Laplace is undoubtedly the most significant publication at the basis of the development of mathematical statistics. Strangely enough, for a period of more than fifty years following the publication of the work of Laplace in 1812, little of importance was contributed to the subject. To be sure, the second law of error of Laplace was developed by Gauss and given its important place in the adjustment of observations, but there was on the whole relatively little progress. Perhaps a complex of causes was involved, but three fairly plausible reasons may be assigned for the lack of contributions to mathematical statistics at this period. First, Laplace left many of his results in the form of approximations

* A Report presented by request of the Program Committee at the symposium held in Chicago, April 25, 1924.

† James Bernoulli, *Ars Conjectandi*, 1713, pp. 210-39 (published eight years after his death).

‡ A. DeMoivre, *Doctrine of Chances* (3rd ed. 1756) pp. 243-54. *Miscellanea Analytica*, 1730, pp. 191-97, Supplement.

§ J. Stirling, *Methodus Differentialis*, 1730, p. 135.

|| C. Maclaurin, *A Treatise on Fluxions*, 1742, p. 672.

¶ L. Euler, COMM. ACAD. PETROP. 6, 1732-33, ed. 1738, pp. 88-97.

** P. S. Laplace, *Théorie Analytique des Probabilités*, 3ième ed., 1820, vol. II, Chap. III, pp. 280-85.

that would not form the basis for further development. Moreover, many of his theorems were not demonstrated with even a fair degree of rigor, and it required the work of Cauchy and others to supply proofs of the theorems. This was important work, but it did not, in general, lead to new results of importance in mathematical statistics. Second, the followers of Gauss promulgated the idea that the deviations from his law of error are due to lack of data, and this attitude was not conducive to the creation of generalized frequency functions. Third, Quetelet* was busy as a popularizer of mathematical statistics. His somewhat sensational language about the stability of social statistics, say of the number of suicides from year to year, caught the imagination; but unfortunately he often asserted the existence of stability on insufficient evidence. The activity of Quetelet cast upon statistics a suspicion of quackery, which still exists to some extent. Moreover, it will probably always be found that those statisticians who use mathematical formulas without guarding well their limitations are likely to have influence productive of an evil very similar to that produced by Quetelet.

An important step in advance was taken in 1877, only three years after the death of Quetelet, in the publication of the theory of Lexis† for the classification of statistical distributions with respect to normal, supernormal, and subnormal dispersion. This theory is based on urn schemata of different constitutions, and pays much attention to the degree of constancy of statistical ratios obtained from different parts of the field of observation. Charlier‡ states that this theory of Lexis is the first essential step forward in mathematical statistics since the days of Laplace. J. M.

---

* A. Quetelet, *Lettres sur la Théorie des Probabilités Appliquée aux sciences, Morales et Politiques*, 1846.

† W. Lexis, *Zur Theorie der Massenerscheinungen in der menschlichen Gesellschaft*, 1877, p. 95.

‡ C. V. L. Charlier, *Vorlesungen über die Grundzüge der mathematischen Statistik*, 1920, p. 5.

Keynes* expresses a somewhat similar view. These may, however, be extreme views, when we take into account the fact that the inequality of Tchebychef was published earlier than the theory of Lexis.

The development of generalized frequency curves and a theory of correlation in the decade 1890 to 1900 started the period of activity in mathematical statistics in which we find ourselves at present.

After a first survey of various topics in the mathematical theory of statistics for consideration in this symposium, it seemed desirable to adopt some principle of selection or elimination of topics. But I have discovered no satisfactory principle of selection except a sort of principle of personal interest which leads me to restrict my remarks to certain points of interest under the following general headings:

I. *Generalized frequency curves.*
II. *Correlation.*
III. *Frequency surfaces.*
IV. *Theory of random sampling.*

No claim is made that the topics selected are more appropriate for consideration in this symposium than others which could easily be named, particularly if we should draw upon recent activities in the special fields of economic,† mortality,‡ and stellar§ statistics. My interest in the general theory of statistics would lead me to include the theory of Lexis with the recent generalization by J. L. Coolidge,|| except for the fact that this theory has become readily accessible to members of the Society.

---

* *A Treatise on Probability*, 1920, p. 393.

† Irving Fisher, *The Making of Index Number*, 1922.

‡ James W. Glover, *U. S. Life Tables*, 1921.
Arne Fisher, *Frequency Curves*, 1922.

§ K. G. Malmquist, *On some relations in stellar statistics*, ARCHIV FÖR MATEMATIK, ASTRONOMI OCH FYSIK, vol. 16, No. 23, 1923.

|| This BULLETIN, vol. 27 (1920–21), p. 439.

I. Generalized Frequency Curves

2. *Introduction.* In the decade from 1890 to 1900, it became well established experimentally that the Gaussian probability function is inadequate to represent all frequency distributions which arise in biological data. When the problem of developing generalized frequency curves was finally attacked, the attack was made from several different directions. Thiele* and Charlier† in Scandinavian countries, Pearson‡ and Edgeworth§ in England, Fechner|| and Bruns¶ in Germany developed theories of generalized frequency curves from viewpoints which give very different degrees of prominence to the Gaussian probability curve in the development of a more general theory. Among other things, I hope to give a brief exposition of the most striking differences in these viewpoints while considering certain properties of the two systems of frequency curves to which I shall direct special attention—the Pearson system and the Charlier system.

3. *The Pearson System of Generalized Frequency Curves.* Pearson's first memoir** dealing with generalized frequency curves appeared in 1895. In this paper he gave four types of frequency curves in addition to the normal curve, with three sub-types under his Type I and two sub-types under his Type III. He published a supplementary memoir†† in

* T. H. Thiele, *Almindelig Iagttagelseslaere*, Copenhagen, 1889.

† C. V. L. Charlier, *Ueber das Fehlergesetz*, Archiv för Matematik, Astronomi och Fysik, vol. 2, No. 8, 1905, pp. 1-9.

——— *Die zweite Form des Fehlergesetzes*, Archiv, vol. 2, No. 15, 1905, pp. 1-8.

‡ Karl Pearson, *Mathematical contributions to the theory of evolution*, Philosophical Transactions, A, vol. 186 (1895), pp. 343-414.

§ F. W. Edgeworth, *The asymmetrical probability-curve.* Philosophical Magazine, vol. 41 (1896), pp. 90-99.

|| G. T. Fechner, *Kollektivmasslehre* (edited by G. R. Lipps), 1897.

¶ H. Bruns, *Ueber die Darstellung von Fehlergesetzen,* Astronomische Nachrichten, vol. 143 (1897).

** Loc. cit., pp. 343-414.

†† Philosophical Transactions, A, vol. 197 (1901), pp. 443-56.

1901 which presented two further types. A second supplementary memoir* which was published in 1916 gave five additional types. Pearson's curves, which are widely different in general appearance, are so well known and so accessible that we shall take no time to comment on them as graduation curves for a great variety of frequency distributions, but we shall attempt to indicate the genesis of the curves with special reference to the methods by which they are grounded on or associated with underlying probabilities.

We shall consider a frequency function $y = F(x)$ of one variable, where $F(x)dx$ differs at most by an infinitesimal of higher order from the probability that x taken at random falls into the interval x to $x+dx$. Pearson's types of curves $y = F(x)$ are obtained by integration of the differential equation

$$\frac{dy}{dx} = \frac{y(x+a)}{c_0 + c_1 x + c_2 x^2}, \tag{1}$$

and by giving attention to the interval on x in which $y = F(x)$ is positive. Obviously, the Gaussian curve is given by the special case $c_1 = c_2 = 0$. We may easily obtain a clear view of the genesis of the system of Pearson curves in relation to laws of probability by following the early steps in the development of equation (1). The development is started by representing the probabilities of successes in n trials given by the terms of the symmetric point binomial $(1/2+1/2)^n$ as ordinates of a frequency polygon at intervals Δx. It is then proved that the slope $\Delta y/\Delta x$ of any side of this polygon is

$$\frac{\Delta y}{\Delta x} = -k^2 y(x+a),$$

where x and y, respectively, are the mean abscissa and the mean ordinate of the side of the polygon. By passing to the limiting situation, we may write

$$\frac{dy}{dx} = -k^2 y(x+a),$$

* PHILOSOPHICAL TRANSACTIONS, A, vol. 216 (1916), pp. 431–57.

from which we obtain the Gaussian curve. The next step consists in dealing with the asymmetric point binomial $(p+q)^n$ in a manner analogous to that used in the case of the symmetric point binomial. This procedure gives the differential equation

$$\frac{dy}{dx} = \frac{y(x+a)}{c_0+c_1x},$$

from which we obtain the Pearson Type III curve

$$y = y_0\left(1+\frac{x}{a}\right)^{\gamma a}e^{-\gamma x}.$$

That is, with respect to the slope property, this curve stands in the same relation to the a priori most probable values given by the asymmetric binomial polygon as the normal curve does to a priori most probable values given by the symmetric binomial. Thus far the underlying probability of success has been assumed constant. The next step consists in taking up a probability problem in which the chance of success is not constant, but depends upon what has happened previously in a set of trials. Thus, the chances of getting $r, r-1, r-2, \ldots, 0$ black balls from a bag containing pn black and qn white balls in drawing r balls one at a time without replacements are given by the successive terms

$$\sum_{s=0}^{s=n}\frac{\binom{r}{s}(qn)_s(pn)_{r-s}}{(n)_r}$$

of a hypergeometric series. When the terms of this series are represented as ordinates of a frequency polygon, and the slope of a side is found in a manner analogous to that used in the case of the point binomial, we obtain the differential equation (1) from which the Pearson curves are obtained by integration.

The idea of obtaining a suitable basis for frequency curves in the probabilities given by terms of a hypergeometric series is the main principle which supports the Pearson curves as statistical probability or frequency curves

rather than as mere graduation curves. That is to say, it is assumed in this system, that the distribution of statistical material may be likened to the law of probability represented by terms of a hypergeometric series. In examining the source of the Pearson curves, the fact should not be overlooked that the Gaussian probability curve can be derived from hypotheses containing much broader implications than are involved in a slope condition of the side of a symmetric binomial polygon. Can the generalized curves likewise be derived from hypotheses involving broader implications than are contained in the slope condition based on the probabilities given by a hypergeometric series? The answer seems to be unknown.

The method of moments plays an essential role in the Pearson system of frequency curves not only in the determination of the parameters but also in providing criteria for selecting the appropriate type of curve. Pearson has attempted to provide a set of curves such that some one of the set would agree with any observational or theoretical frequency curve of positive ordinates to the extent of having equal areas and equal first, second, third, and fourth moments of area about a centroidal axis.

Let μ_s be the sth moment coefficient about a centroid vertical taken as the y-axis. That is, let

$$\mu_s = \int_{-\infty}^{\infty} F(x) x^s dx.$$

Next, let $\beta_1 = \mu_3^2/\mu_2^3$ and $\beta_2 = \mu_4/\mu_2^2$. Then it is Pearson's thesis that the conditions $\mu_0 = 1$, $\mu_1 = 0$ together with the equality of the numbers μ_2, β_1, and β_2 for the observed and theoretical curves lead to equations whose solutions give such values to the parameters of the frequency function that we almost invariably obtain excellency of fit by using the appropriate one of the curves of his system to fit the data, and that badness of fit can be traced, in general, to heterogeneity of data, or to the difficulty in the determination of moments from the data as in the case of J and U shaped curves.

Let us next examine the nature of the criteria by which to pass judgment on the type of curve to use in any numerical case. Obviously, the form which the integral $y = F(x)$ obtained from (1) takes, depends on the nature of the zeros of the quadratic function in the denominator. An examination of the discriminant of this quadratic function leads to equalities and inequalities involving β_1 and β_2 which serve as criteria in the selection of the type of function to be used. A systematic procedure for applying criteria has been thoroughly developed.* The relations between β_1 and β_2 are represented by curves in the $\beta_1\beta_2$-plane. To illustrate, the normal curve corresponds to the point $\beta_1 = 0$, $\beta_2 = 3$ in this plane. Type III is to be chosen when the point (β_1, β_2) is on the line $2\beta_2 - 3\beta_1 - 6 = 0$; and Type V, when (β_1, β_2) is on the cubic

$$\beta_1(\beta_2 + 3)^2 = 4(4\beta_2 - 3\beta_1)(2\beta_2 - 3\beta_1 - 6).$$

In considering sub-types under Type I, the biquadratic

$$\beta_1(8\beta_2 - 9\beta_1 - 12)(\beta_2 + 3)^2 = (10\beta_2 - 12\beta_1 - 18)^2(4\beta_2 - 3\beta_1)$$

separates the area of J curves or modeless curves from the area of limited range modal curves and the area of U curves. Without going into further detail about criteria for the selection of the type of curve, we may summarize by saying that curves traced on the $\beta_1\beta_2$-plane provide the means of selecting the Pearson type of frequency curve appropriate to the given distribution in so far as the necessary conditions expressed by relations between β_1 and β_2 turn out to be sufficient to determine a suitable type of curve.

4. *Generalized Normal Curve.—Charlier System.* As indicated in § 3, the Pearson system of frequency curves assigns only a *point* in the $\beta_1\beta_2$-plane for the region of applications of the Gaussian law in fitting frequency distributions. It seems not unnatural, however, to have some

* A. Rhind, BIOMETRIKA, vol. 7 (1909–10), pp. 127–35; cf. *Tables for Statisticians and Biometricians*, 1914, pp. ix–ixx, and pp. 66–67; see also, Karl Pearson, loc. cit., PHILOSOPHICAL TRANSACTIONS, A, vol. 216 (1916), pp. 429–57.

doubt about the Gaussian curve taking such a small place in the representation of frequency, and to turn with considerable interest to the Charlier system of representation of frequency which gives great prominence to the Gaussian probability function

$$y = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-b)^2}{2\sigma^2}}.$$

Among the early contributors to the theory of the generalized normal curve we find Gram,* Thiele,† Edgeworth,‡ Bruns§ and Charlier.|| Of the various contributions to the subject, those of Charlier are particularly elegant and noteworthy. Charlier has shown by extensions of the Laplace theory based on the hypothesis of elementary errors that the law of error assumes one of the following two forms:

TYPE A.

$$F(x) = a_0\Phi(x) + a_3\Phi^{(3)}(x) + a_4\Phi^{(4)}(x) + \cdots + a_n\Phi^{(n)}(x) + \cdots,$$

where

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}},$$

and $\Phi^{(n)}(x)$ is the nth derivative of $\Phi(x)$ with respect to x.

TYPE B.

$$F(x) = c_0\psi(x) + c_1\Delta\psi(x) + c_2\Delta^2\psi(x) + \cdots + c_n\Delta^n\psi(x) + \cdots,$$

where

$$\psi(x) = e^{-\lambda}\frac{\sin\pi x}{\pi}\left\{\frac{1}{x} - \frac{\lambda}{(x-1)1!} + \frac{\lambda^2}{(x-2)2!} - \frac{\lambda^3}{(x-3)3!} + \cdots\right\} = \frac{e^{-\lambda}\lambda^x}{x!},$$

the Poisson exponential for non-negative integral values of x.

* J.P. Gram, *On Raekkeudvidlinger, bestemte ved Hjaelp af de minste Kradvaters Methode* (Doctor's dissertation), Copenhagen, 1879.

† T. N. Thiele, *Almindlig Iagttagelseslaere*, 1889; cf. Thiele, *Theory of Observations*, 1903.

‡ F. Y. Edgeworth, loc. cit; also *The law of error*, CAMBRIDGE PHILOSOPHICAL TRANSACTIONS, vol. 20 (1904), pp. 36–65, 113–41.

§ H. Bruns, loc. cit; also *Wahrscheinlichkeitsrechnung und Kellektivmasslehre*, 1906.

|| C. V. L. Charlier, loc. cit; also *Ueber Darstellung willkürlicher Functionen*, ARCHIV FÖR MATEMATIK, ASTRONOMI OCH FYSIK, vol. 2, No. 20, 1905, pp. 1–35.

The Type A series represents an arbitrary function $F(x)$ subject to certain conditions of continuity* and vanishing at infinity. The coefficients a_n in Type A may be expressed in terms of moments of area under the given frequency curve because the functions $\Phi^{(n)}(x)$ and the Hermite polynomials $H_n(x)$ defined by the equation

$$\Phi^{(n)}(x) = (-1)^n H_n(x)\Phi(x)$$

form a biorthogonal system. Thus

$$a_n = \frac{(-1)^n \int_{-\infty}^{\infty} F(x) H_n(x)\, dx}{n!}.$$

Since $H_n(x)$ is a polynomial of degree n in x, the coefficients a_n are thus given in terms of moments of area under the frequency curve. Moreover, the value of any coefficient thus obtained is the same as that obtained by finding the best approximation to $F(x)$ in the sense of a certain least squares criterion by the first s terms of the series ($s > n$). In the Type B series, the coefficients may be determined by the method of moments or by means of the semi-invariants† of Thiele.

To be of practical value in fitting given numerical distributions, it is essential that only a few terms of the series in Type A be required to fit the distribution. The closeness with which the first few terms will represent a given function $F(x)$ depends much on the extent to which the generating function $\Phi(x)$ is a fair approximation to the distribution. In case the generating function $\Phi(x)$ is not even a rough approximation to the distribution, it may be possible to introduce in place of $\Phi(x)$ a function of approximation $\Theta(x)$ as a generating function in the series. N. R. Jørgensen‡ has used the function

* Wera Myller-Lebedeff, MATHEMATISCHE ANNALEN, vol. 64 (1907), p. 338; and H. Weyl, MATHEMATISCHE ANNALEN, vol. 66 (1908), p. 306.

† Arne Fisher, *The Mathematical Theory of Probabilities,* 1922. p. 271.

‡ *Undersögelser over Frequensflader og Korrelation,* 1916, pp. 177-93.

$$\Theta(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(\log x - m)^2}{2\sigma^2}}$$

as a generating function and has by this means succeeded in representing closely some remarkably skew distributions. H. C. Carver* and Emeterio Roa have done some work on this plan of representation by introducing various types of generating functions.

With respect to the selection of Type A or Type B of Charlier to represent given data, no criterion has been given which enables one to distinguish sharply between cases in which to apply one of these types in preference to the other, but Type B applies in general, to decidedly skew distributions, and particularly to those defined only at integral values of the variable when the Poisson exponential is the generating function. While the systematic procedure in fitting Charlier curves to data is thus not so well standardized as the methods used in fitting curves of the Pearson system to data, tables of $\Phi(t)$, where t is in units of standard deviation, of its integral $\int_0^t \Phi(t)dt$, and of its second to eighth derivatives are given to five decimal places for $t = 0$ to $t = 5$ at intervals of .01 by James W. Glover,† and tables of the function, its integral and first six derivatives are given by N. R. Jørgensen‡ to seven decimal places for $t = 0$ to $t = 4$.

The question naturally arises as to the arguments which support the Charlier system of representation in comparison with the Pearson system. The Charlier system is surely the better grounded in the theory of probability, and is adapted to the representation of an arbitrary function subject to reasonable conditions of continuity and vanishing at infinity. A disadvantage of the system is found in the fact that its application to numerical distributions is likely to be laborious, and the probable errors of the coefficients of

* Cited in *Handbook of Mathematical Statistics,* by H. L. Rietz and others, 1924, p. 116.

† *Tables of Applied Mathematics*, 1923, pp. 392-411.

‡ Loc. cit., p. 178-93.

the series are large if we find it necessary to use more than three or four significant terms. But even if the representation by the Charlier system proves to be so laborious that it is rarely used with the more common numerical distributions, the series is nevertheless of great value in the representation of laws of probability. To illustrate, take the problem of the distribution of the sum of n numbers selected at random from a uniform distribution. Laplace gave an approximate solution of the problem adapted to computation for the case when n is large, and applied the result of his theory to the question of the random distribution of the orbits of comets. The method of Laplace in obtaining the approximation is of doubtful validity. Cauchy* put Laplace's approximation on a much more rigorous basis in a memoir published in 1841. In a paper which the writer presented to the Society last April, it is shown that the representation is given by the Type A function of Charlier, and that each additional term improves the approximation in the sense of a certain least squares criterion. Again, B. H. Camp in a recent paper† on certain important problems in sampling has made much use of the Type A representation in his theory. The point in citing these illustrations is that the Charlier representation is likely to be found very useful in the general theory of probability, apart from the fitting of frequency curves to numerical data.

5. *Transformation of Frequency Functions.* Before leaving the subject of frequency functions of one variable, let us consider briefly the idea of regarding certain frequency functions as the result of transformation of the independent variable in simple frequency functions. F. Y. Edgeworth‡

* A. L. Cauchy, JOURNAL DE L'ECOLE POLYTÉCHNIQUE, Cahier 28, vol. 21 (1841), pp. 147-248.

† JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION, vol. 18 (1923), pp. 964-77.

‡ JOURNAL OF THE ROYAL STATISTICAL SOCIETY, vol. 61 (1898), pp. 670-700.

and J. C. Kapteyn* proposed that skew frequency curves should be regarded as transformations of the Gaussian curve. It seems that this idea would accord with what happens in certain natural phenomena, although it is difficult to predict its generality. That is to say, if certain values are distributed normally, we inquire into the distribution of certain simple functions of these values. For example, the writer gave in a recent paper† certain properties of frequency curves obtained when the values of a normally distributed variable, x, are transformed by

$$x = kx'^n.$$

It is shown that the form and properties of the resulting distribution differ widely according as $n < 0$, $0 < n < 1$, and $n > 1$. As another example, N. R. Jørgensen has made $x = \log x'$ in the Gaussian frequency function thus obtaining

$$y = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(\log x' - m)^2}{2\sigma^2}}.$$

He found the semi-invariants of Thiele for this function and showed that the function may replace the Gaussian function as a generating function in the Charlier series of Type A for the representation of certain skew distributions. Thus, it seems that functions arising from the transformation of variables may be found to be useful.

II. Correlation

6. *Simple Correlation.* It seems that Francis Galton‡ was the first to deal with correlation among direct observations. If we should follow the historical development of correlation, we should simply give Galton's definition of correlation and his ideas of the regression of one variable on another before proceeding to correlation surfaces in three dimensions. For example, the mathematical solution of the special correlation problem proposed by Francis

* *Skew Frequency Curves in Biology and Statistics,* Groningen, 1903.

† H. L. Rietz, Annals of Mathematics, vol. 23 (1922), pp. 292-300.

‡ Proceedings of the Royal Society, vol. 40 (1886), p. 42.

Galton to J. D. Hamilton Dickson in 1886* consisted simply in giving the equation of a normal frequency surface to correspond to given standard deviations and regression lines. Furthermore, the early contributions of Karl Pearson to correlation theory involving the influence of selection were concerned with frequency surfaces.† But, beginning with a paper by G. Udny Yule in 1897, the theory of correlation has been developed without limitation to a particular type of frequency surface. It is of some interest that Yule returned very close to the primary ideas of Galton, by placing the emphasis on the lines of regression. This method of Yule may be appropriately called the regression method of approaching correlation in contrast to the frequency surface method. It is our purpose to present enough of the elements of the regression method to give a basis for a brief exposition of the nature of certain recent contributions to correlation from the regression standpoint. Special attention will be directed (1) to the general method of determining the successive terms of a non-linear regression equation, (2) to the connection of the correlation coefficient and regression curves with some simple problems of a priori probabilities, (3) to the development of a theory of correlation of n variables in the case of non-linear regression by means of multiple and partial correlation ratios.

To introduce a convenient notation let $(X_1, Y_1), (X_2, Y_2), \ldots, (X_N, Y_N)$ be pairs of real numbers such that at least two of the X's are unequal and at least two of the given Y's are unequal. Let $\overline{X}, \overline{Y}$ be the arithmetic means of the given values of X's and Y's respectively.

Then let

$$x_i = \frac{X_i - \overline{X}}{\sigma_x}, \qquad y_i = \frac{Y_i - \overline{Y}}{\sigma_y},$$

where σ_x and σ_y are respectively standard deviations (root-

* PROCEEDINGS OF THE ROYAL SOCIETY, vol. 40 (1886), p. 63.

† PHILOSOPHICAL TRANSACTIONS, vol. 187 (1896), pp. 253–318.

mean-squares) of X's and Y's, so that each deviation is expressed by its ratio to the standard deviation of the system to which it belongs. Then the correlation coefficient r is defined by

$$r = \frac{1}{N}\sum_{1}^{N} x_i y_i. \tag{1}$$

By writing (1) in the form

$$r = 1 - \frac{1}{2N}\sum_{1}^{N} (x_i - y_i)^2, \tag{2}$$

and

$$r = -1 + \frac{1}{2N}\sum_{1}^{N} (x_i + y_i)^2, \tag{3}$$

it follows at once that $-1 \leqq r \leqq 1$, and that the X_i's and Y_i's are linearly related when $r = 1$ or -1. In papers by Huntington* and Jackson,† the significance of r is brought out in interesting ways by interpretations of $\sum_{1}^{N} (x_i - y_i)^2/(2N)$ in (2).

In the sense of a certain least squares criterion, we may obtain the values of y corresponding to assigned values of x more accurately from $y = rx$ than from any other linear equation. The line $y = rx$ is called the line of regression of y on x. The mean square of the errors involved in using values of $y = rx$ for the given y's is $s_y^2 = 1 - r^2$. Thus,

$$r^2 = 1 - s_y^2. \tag{4}$$

Let us now conceive of dividing the whole interval along the x-axis which includes our data into suitable equal class intervals Δx. Then the y's which correspond to the x's in the interval Δx are called an x-array of y's. When the means of the arrays of y's are on the line of regression $y = rx$, the regression of y on x is said to be linear. When

* E. V. Huntington, AMERICAN MATHEMATICAL MONTHLY, vol. 26 (1919), p. 425.

† Dunham Jackson, AMERICAN MATHEMATICAL MONTHLY, vol. 31 (1924), p. 117.

the means of the arrays of y's are far from the line of regression, the value of the correlation coefficient is likely to be misleading. For example, we* may have $r = 0$ for variables x and y when y is a simple periodic function of x. To characterize correlation in such situations, Karl Pearson devised† a measure of correlation called the *correlation ratio,* which we shall now describe briefly. By analogy with (4), we define

$$\eta_{yx}^2 = 1 - s_y'^2 \tag{5}$$

as the correlation ratio of y on x, where $s_y'^2$ is the mean square of deviations of y's from the *means* of arrays, and these means of arrays need not lie on or near a straight line. The s_y' in (5) agrees with the s_y in (4) only when the regression is linear and $\Delta x \to 0$. It is obvious that when s_y' is small in comparison to unity there is a tendency for the points of the scatter diagram to concentrate in a narrow band along the regression curve, and we have a high degree of correlation. It is an easy step from (5) to deduce

$$\eta_{yx} = \sigma_{\bar{y}_x}, \tag{6}$$

where $\sigma_{\bar{y}_x}$ is the standard deviation of means of x-arrays of Y's when the square of each deviation of a mean of an array is weighted with the number in the array. It is easily shown that

$$1 \geqq \eta_{yx}^2 \geqq r^2,$$

and that the equality $\eta_{yx}^2 = r^2$ holds only in the case of linear regression. It may be noted from (6) that the correlation ratio of y on x is the ratio $\sigma_{\bar{y}_x}/\sigma_y$, if the unit for measuring values of y is not σ_y.

As early as 1905, the parameters of the special regression curves given by polynomials $y = f(x)$ of the second and

* See H. L. Rietz, QUARTERLY PUBLICATION, AMERICAN STATISTICAL ASSOCIATION, vol. 16 (1919), pp. 472–76.

† *On the general theory of skew correlation and non-linear regression,* DRAPERS' COMPANY RESEARCH MEMOIRS, BIOMETRIC SERIES II (1905), pp. 1–54.

third degrees were determined in terms of power moments and product moments. In 1921, Karl Pearson* published a general method of determining successive terms of the regression curve of the form

$$y = f(x) = a_0\psi_0 + a_1\psi_1 + \cdots + a_n\psi_n, \tag{7}$$

where $a_0, a_1, \ldots, a_n$ are constants to be determined and ψ_s is an orthogonal function of x. That is,

$$\sum(N_x\psi_s\psi_{s'}) = 0, \qquad (s \neq s'),$$

if the summation $\sum$ be taken for all values of x corresponding to an arbitrary system of arrays with frequency in an x-array given by N_x.

7. *Simple Correlation and Probability.* Thus far in this lecture correlation has been discussed by means of averages, ratios of averages, and by the correspondence between an assigned value of one variable and an average value of another. Probability theory has not entered in explicit form. Before leaving simple correlation, I wish to say that it has seemed important to me to construct urn schemata which would give a meaning to the correlation coefficient in pure chance. In a paper† published in 1920, certain urn schemata were devised which give linear regression and very simple values for the correlation coefficient. Other schemata apparently equally simple give non-linear regression. The general plan of the schemata consisted in requiring certain elements to be common in successive random drawings. By means of partial correlation coefficients, J. R. Musselman‡ recently gave simple and interesting proofs of values of the correlation coefficients for those of my urn schemata in which the regression is linear. His method does not, however, replace my method because he assumes the existence of linear regression, which is proved in my

* BIOMETRIKA, vol. 13 (1921), p. 296.

† H. L. Rietz, ANNALS OF MATHEMATICS, vol. 21 (1920), pp. 307–22.

‡ JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION, vol. 18 (1923), pp. 908–11.

paper. It is hoped that further contributions to correlation involving urn schemata will give the theory of correlation a more prominent place in the pure theory of probability.

8. *Multiple Correlation.* Given N sets of values* of n real variables $x_1, x_2, \ldots, x_n$, where any variable, x_j, is referred to the arithmetic mean of its N given values as an origin, and is measured in units of the standard deviation, σ_j, of its N given values. Let r_{pq} be the correlation coefficient of the N given values of x_p and x_q. Then we seek to determine the parameters in the linear regression equation

$$x = b_{12}x_2 + b_{13}x_3 + \cdots + b_{1n}x_n + c \tag{8}$$

so that x computed from (8) will give the "best" estimates of the values of x_1 to correspond to assigned values of $x_2, x_3, \ldots, x_n$.

Adopting a least squares criterion,† we determine the coefficients in (8) so that

$$(x_1 - b_{12}x_2 - b_{13}x_3 - \cdots - b_{1n}x_n - c)^2$$

shall be a minimum. This gives for the regression equation of x_1 on $x_2, x_3, \ldots, x_n$

$$x = -\sum_{p=2}^{p=n} \frac{R_{p1}}{R_{11}} x_p, \ldots, \tag{9}$$

where R_{pq} is the cofactor of the pth row and qth column of the determinant

$$R = \begin{vmatrix} 1 & r_{12} & r_{13} & \cdots & r_{1n} \\ r_{21} & 1 & r_{23} & \cdots & r_{2n} \\ r_{31} & r_{32} & 1 & \cdots & r_{3n} \\ \cdot\cdot & \cdot\cdot & \cdot\cdot & \cdots & \cdot\cdot \\ \cdot\cdot & \cdot\cdot & \cdot\cdot & \cdots & \cdot\cdot \\ r_{n1} & r_{n2} & \cdot\cdot & \cdots & 1 \end{vmatrix}. \tag{10}$$

The correlation coefficient $r_{1 \cdot 23 \ldots n}$ between the observed values of x_1 and its corresponding estimated values x

* We assume that at least two of the N given values of a variable x_1 are unequal.

† G. Udny Yule, JOURNAL OF THE ROYAL STATISTICAL SOCIETY, vol. 60 (1897), p. 812.

calculated from the linear function (9) of $x_2, x_3, \ldots, x_n$ is called the multiple correlation coefficient of x_1 with the other $n-1$ variables. The multiple correlation coefficient $r_{1 \cdot 23 \ldots n}$ is expressible in terms of simple correlation coefficients by the formula

$$r_{1 \cdot 23 \ldots n} = \sqrt{1 - \frac{R}{R_{11}}}. \tag{11}$$

If the scatter $\sigma_{1 \cdot 23 \ldots n}$ of the observed values of x_1 from the regression hyperplane (9) is defined as the square root of mean square, that is,

$$\sigma^2_{1 \cdot 23 \ldots n} = \frac{\sum (x_1 - x)^2}{N},$$

it can be proved that

$$\sigma_{1 \cdot 23 \ldots n} = \sqrt{\frac{R}{R_{11}}}, \tag{12}$$

and from (11) and (12) that

$$\sigma^2_{1 \cdot 23 \ldots n} = 1 - r^2_{1 \cdot 23 \ldots n}. \tag{13}$$

9. *Partial Correlation.* It is often important to obtain the degree of correlation between two variables x_1 and x_2 when the other variables $x_3, x_4, \ldots, x_n$ have assigned values. To illustrate, we may have found a correlation between characteristics A and B. A plausible interpretation may be that the correlation thus found is due to the correlation of each of them with C. In this case we could remove the influence of C if we have an unlimited amount of data by restricting our data to a universe of A and B corresponding to an assigned C. But usually the data are not readily available for such a procedure. Instead of thus restricting data, we would often make use of a partial correlation coefficient. We define $r_{12 \cdot 34 \ldots n}$ as the partial correlation coefficient between x_1 and x_2 when we have eliminated the influences of the variables $x_3, x_4, \ldots, x_n$ in so far as they can be eliminated by means of a linear function of these variables. That is, the partial correlation $r_{12 \cdot 34 \ldots n}$ is the correlation between residuals

$$x_{1\cdot 34\ldots n} = x_1 - b_{13}\cdot x_3 - \cdots - b_{1n}\cdot x_n,$$

and

$$x_{2\cdot 34\ldots n} = x_2 - b_{23}\cdot x_3 - \cdots - b_{2n}\cdot x_n,$$

in estimating x_1 and x_2 by means of linear functions of $x_3, x_4, \ldots, x_n$ and this partial correlation is said to be of order $n-2$, the number of variables held constant. It would ordinarily involve a large amount of labor to apply this definition directly to data. Fortunately the partial correlation coefficient is expressible directly in terms of simple correlation coefficients by the formula

$$r_{12\cdot 34\ldots n} = \frac{-R_{12}}{\sqrt{R_{11}R_{22}}}. \tag{14}$$

An important relation between partial and multiple correlation coefficients may now be derived. From (11) and (14),

$$1 - r^2_{1\cdot 23\ldots n} = \frac{R}{R_{11}}, \qquad r_{12\cdot 34\ldots n} = \frac{R_{12}}{\sqrt{R_{11}R_{22}}}.$$

Hence we have

$$1 - r^2_{12\cdot 34\ldots n} = \frac{R_{11}R_{22} - R_{12}^2}{R_{11}R_{22}}.$$

By a well known theorem of determinants,*

$$\begin{vmatrix} R_{11} & R_{12} \\ R_{12} & R_{22} \end{vmatrix} = R_{11}R_{22} - R_{12}^2 = R\,R_{11\,22}.$$

Hence we have

$$1 - r^2_{12\cdot 34\ldots n} = \frac{R\cdot R_{11\,22}}{R_{11}R_{22}} = \frac{\dfrac{R}{R_{11}}}{\dfrac{R_{22}}{R_{11\,22}}} = \frac{1 - r^2_{1\cdot 23\ldots n}}{1 - r^2_{1\cdot 34\ldots n}}. \tag{15}$$

Thus we can express the partial correlation coefficient $r_{12\cdot 34\ldots n}$ of order $n-2$ in terms of the multiple correlation coefficient $r_{1\cdot 23\ldots n}$ of order $n-1$ and the multiple correlation coefficient $r_{1\cdot 34\ldots n}$ of order $n-2$.

* Maxime Bôcher, *Introduction to Higher Algebra,* 1912, p. 33.

10. *Non-linear Regression in n Variables.—Multiple Correlation Ratio.* L. Isserlis* and Karl Pearson† have developed a theory of non-linear regression in the case of more than two variables. Consider the variables $x_1, x_2, x_3, \ldots, x_n$, and we have an array of observed values of x_1 whose mean value $\bar{x}_{1\cdot 23\ldots n}$ we may appropriately call the partial mean value of x's for constant $x_2, x_3, \ldots, x_n$. Then we may define the multiple correlation ratio $\eta_{1\cdot 23\ldots n}$ of x_1 on $x_2, x_3, \ldots, x_n$ by writing

$$\eta^2_{1\cdot 23\ldots n} = \frac{\sum_{x_2}\sum_{x_3}\cdots\sum_{x_n}(N_{23\ldots n}\bar{x}^2_{1\cdot 23\ldots n})}{N}. \tag{16}$$

$$= \sigma^2_{\bar{x}_{1\cdot 23\ldots n}} \tag{17}$$

where $\sigma_{\bar{x}_{1\cdot 23\ldots n}}$ is the standard deviation of the means of arrays computed by weighting the squares of the deviations of these means from the mean $\bar{x}_1 = 0$ of all x's with the number $N_{23\ldots n}$ in the array. It may be observed from (17) that $\eta_{1\cdot 23\ldots n}$ is the ratio $\sigma_{\bar{x}_{1\cdot 23\ldots n}}/\sigma_1$, if the unit for measuring values of x_1 is not σ_1. We now define $\sigma'_{1\cdot 23\ldots n}$ by the equation

$$\sigma'^2_{1\cdot 23\ldots n} = \frac{\sum_{x_2}\sum_{x_3}\cdots\sum_{x_n}\{N_{23\ldots n}(x_1-\bar{x}_{1\cdot 23\ldots n})^2\}}{N}, \tag{18}$$

the mean‡ square of standard deviations of arrays of x_1 for assigned values of other $n-1$ variables $x_2, x_3, \ldots, x_n$. From (16) and (18), we find

$$\sigma'^2_{1\cdot 23\ldots n} = 1-\eta^2_{1\cdot 23\ldots n}. \tag{19}$$

It may be recalled that our definition of η_{12} for two variables is such that $1-\eta^2_{12}$ is the mean square of standard deviations of the arrays of x_1 which correspond to assigned values of x_2. Next consider n variables, and let x_1 be

* BIOMETRIKA, vol. 10 (1914–15), pp. 393-411.

† PROCEEDINGS OF THE ROYAL SOCIETY, A, vol. 91 (1915), pp. 492-98.

‡ The square of each standard deviation is weighted with the number in the array.

limited to those $N_{34\ldots n}$ values for which $x_3, x_4, \ldots, x_n$ are assigned. Then consider the expression

$$\sigma'^2_{1.34\ldots n}(1-\eta^2_{12.34\ldots n}), \tag{20}$$

where $\sigma'^2_{1.34\ldots n}$ is the mean square of standard deviations of arrays of x_1 for assigned values of $x_3, x_4, \ldots, x_n$. In this restricted universe, we deal with two variables, x_1 and x_2, and (20) is the mean square deviation for assigned values of $x_2, x_3, x_4, \ldots, x_n$, where $\eta_{12.34\ldots n}$ is the partial correlation ratio of x_1 on x_2 for constant $x_3, x_4, \ldots, x_n$. That is

$$\sigma'^2_{1.34\ldots n}(1-\eta^2_{12.34\ldots n}) = 1-\eta^2_{1.23\ldots n}. \tag{21}$$

Analogous to (19), we may write

$$\sigma'^2_{1.34\ldots n} = 1-\eta^2_{1.34\ldots n}. \tag{22}$$

Hence from (21) and (22), we have

$$1-\eta^2_{12.34\ldots n} = \frac{1-\eta^2_{1.23\ldots n}}{1-\eta^2_{1.34\ldots n}}. \tag{23}$$

From (23), we note that the partial correlation ratio of order $n-2$ can be expressed in terms of multiple correlations of order $n-1$ and $n-2$ in a form exactly analogous to that for expressing partial correlation coefficients in terms of multiple correlation coefficients. While the method of computing $\eta_{1.23\ldots n}$ is simple in principle, it is unfortunately laborious from the arithmetic standpoint. It is important as a next step in the investigation to discover a way of expressing multiple correlation ratios in terms of simple correlation ratios just as we know how to express multiple correlation coefficients in terms of simple correlation coefficients. L. Isserlis has taken a step in this direction by showing that, for a certain type of quadric regression surface, the direct calculation of the multiple correlation ratio may be replaced by the calculation of four simple correlation ratios. But the general problem of expressing the multiple correlation ratio in terms of simple correlation ratios is still unsolved.

III. Frequency Functions of n Variables—Correlation Surfaces

11. *Normal Correlation Surfaces.* The function

$$z = f(x_1, x_2, \ldots, x_n)$$

is called a frequency* function of the n variables $x_1, x_2, \ldots, x_n$ if

$$z\, dx_1\, dx_2 \cdots dx_n$$

gives to within infinitesimals of higher order the probability that a set values of $x_1, x_2, \ldots, x_n$ taken at random will fall respectively into the intervals x_1 to $x_1 + dx_1$, x_2 to $x_2 + dx_2, \ldots,$ x_n to $x_n + dx_n$. With the notation of § 7 on multiple correlation, the natural extension of the Gaussian frequency function of one variable to the case of n normally correlated variables $x_1, x_2, \ldots, x_n$ gives a frequency function of the exponential type

$$z = z_0\, e^{-\frac{1}{2}\Phi}, \tag{1}$$

where Φ is a homogeneous quadratic function of the n variables and may be written in the form

$$\Phi = \frac{1}{R}\left(R_{11}x_1^2 + R_{22}x_2^2 + \cdots + 2R_{12}x_1x_2 + \cdots\right), \tag{2}$$

the determinant R with correlation coefficients as elements and its cofactors R_{pp} and R_{pq} being defined in § 7.

Karl Pearson† published the general equation of this frequency surface in $n+1$ dimensions and dealt with certain of its important properties in 1896. F. Y. Edgeworth‡ had partially developed the theory of this surface as early as 1892. In three and four dimensions the form of the surface dates back to Bravais§ in 1846, but not as a surface of distribution of directly observed statistical measurements.

* Charlier calls $f(x_1, x_2, \ldots, x_n)$ a correlation function. See Archiv för Matematik, Astronomi och Fysik, vol. 8, No. 4 (1912).

† Philosophical Transactions, A, vol. 187 (1896), pp. 253-318.

‡ Philosophical Magazine, (5), vol. 34 (1892), pp. 190-204.

§ *Sur les probabilités des erreurs de situation d'un point*, Mémoires par divers Savants, vol. 9 (1846), pp. 255-332.

Bravais considered the distribution of linear functions of independent errors in observed quantities rather than directly observed correlated variables. J. L. Coolidge* recently derived the Gaussian law of error for n variables by stating explicitly a set of underlying assumptions. He obtained a surface of the form (1), and determined the parameters by expressing moments of actual errors in terms of moments of residuals.

In our notation for simple correlation, § 6, the surface (1) in three dimensions takes the well known form

$$z = \frac{1}{2\pi\sqrt{1-r^2}}\, e^{-\frac{1}{2(1-r^2)}(x^2+y^2-2rxy)}. \tag{3}$$

The equal-frequency curves obtained by making z take constant values are an infinite system of homothetic ellipses, any one of which has an equation of the form

$$x^2+y^2-2rxy = \lambda^2. \tag{4}$$

When these ellipses are represented on the xy-plane (scatter-diagram plane), the probability that an (x, y) taken at random will lie within the ellipse (4) is given by

$$1-e^{-\frac{\lambda^2}{2(1-r^2)}}. \tag{5}$$

The particular ellipse of the system such that the probability that an (x, y) taken at random will fall within it is one-half, is called the *probable ellipse* and has been frequently discussed. In a paper published in 1912, the writer† defined the *ellipse of maximum probability* as that ellipse of the system along which, for a given small ring $d\lambda$, we expect a greater frequency than along any other ellipse of the system. This ellipse is given by making $\lambda^2 = 1-r^2$ in (4). It is a fact of some interest that this ellipse is the locus of parabolic points of the correlation surface.

One of the most interesting problems I have studied in connection with this surface relates to the determination of

* TRANSACTIONS OF THIS SOCIETY, vol. 24 (1922), pp. 135-43.

† H. L. Rietz, ANNALS OF MATHEMATICS, vol. 13 (1912), pp. 187-99.

the locus along which the frequency or density of points on the plane of distribution (scatter diagram) bears a simple relation to the corresponding density under independence. Thus, we seek the curve along which dots of the scatter diagram are k times as frequent as they would be under independence where k is a constant. Equating z in (3) to k times the corresponding value of z when $r = 0$ in (3), we obtain the hyperbola

$$x^2 + y^2 - \frac{2}{r}xy = \frac{(1-r^2)}{-r^2}\log(k^2 - k^2r^2). \tag{6}$$

Karl Pearson had dealt* with the particular case of this curve for $k = 1$. I was impressed by the fact that the density of distribution at the centroid in (3) is $1/\sqrt{1-r^2}$ times as much as it would be under independence and was naturally inclined to inquire about the locus of all points for which $k = 1/\sqrt{1-r^2}$ in (6). It turns out that in this case the hyperbola degenerates into straight lines

$$y = \frac{1}{r}x(1 \pm \sqrt{1-r^2}). \tag{7}$$

These lines separate the plane of distribution into four compartments such that 1/4 is the probability that a pair of values (x, y) taken at random will give a point falling into any prescribed one of these compartments. Although no further discussion of the properties of normal correlation surfaces will be attempted in this paper, certain properties analogous to those mentioned for the surface in three dimensions would probably follow rather readily in the case of the surfaces in higher dimensions. The system of ellipsoids of equal frequencies has been studied to some extent.† In a recent paper by James McMahon,‡ the connection between the geometry of the hypersphere and the theory of normal

* Drapers' Company Research Memoirs, Biometric Series I, vol. 13, p. 10.

† See E. Czuber, *Theorie der Beobachtungsfehler*, 1891, pp. 355-82.

‡ Biometrika, vol. 15 (1923), pp. 192-208, paper edited by F. W. Owens after the death of Professor McMahon.

frequency functions of n variables is established by linearly transforming the hyperellipsoids of equal frequency into a family of hyperspherical surfaces, and by applying the formulas of hyperspherical goniometry to obtain theorems in multiple and partial correlations.

12. *Generalized Frequency Surfaces.* The history of the difficulties which have been encountered in efforts to reach general skew frequency surfaces has been given recently by Karl Pearson.* He tells us that in 1895 after the publication of his memoir on skew frequency curves, he proceeded to the problem of skew frequency surfaces, and gave much time to attempting its solution. He became convinced on experimental grounds that a generalized frequency surface could not be obtained by taking the product of two of his skew frequency functions and transforming coordinates. Pearson next approached the problem by endeavouring to determine surfaces which should grow out of the double hypergeometric series in a way analogous to that by which his skew frequency curves arise from the single hypergeometric series. He obtained the differential equations from the series, but he tells us he has failed to integrate them although he has returned to them again and again for nearly thirty years. In 1901, Pearson put the problem before L. N. G. Filon, who made some progress by obtaining certain special surfaces. Pearson had also obtained a special surface for linear regression and for what he calls parabolic variance. In 1914, Pearson put his differential equations before L. Isserlis. In a paper on the application of *Solid hypergeometric series to frequency distributions of space*†, Isserlis solved the problem of fitting a double hypergeometric series to certain distributions of two variables.

In 1923, Seimatsu Narumi—a Japanese mathematician—published‡ an important contribution to the solution of the

* BIOMETRIKA, vol. 15 (1923), p. 222.

† PHILOSOPHICAL MAGAZINE, (6), vol. 28 (1914), p. 279.

‡ BIOMETRIKA, vol. 15 (1923), pp. 77–88; pp. 208–221.

problem. It appears that Pearson suggested* to Narumi the problem of working from functional equations, with assumed forms of regression and scedastic functions, back to the frequency surface. All I shall attempt here is to give a general idea of this method of approach, and to state a few of the most interesting results. Let the regression curve of y on x be $y = f_2(x)$, and that of x on y be $x = f_1(y)$. Narumi gives

$$(8) \qquad \begin{aligned} z &= \Phi_1(y)\,\Psi_1[\{x - f_1(y)\}F_1(y)] \\ &= \Phi_2(x)\,\Psi_2[\{y - f_2(x)\}F_2(x)], \end{aligned}$$

as the general functional equation of the frequency surfaces. One way of regarding this equation is to consider $\Phi_1(y)$ as giving the relative frequency of values existing for a total array curve corresponding to an assigned y. With the array curve assigned, $\Psi_1[\{x - f_1(y)\}F_1(y)]$ would give relative frequencies at any point along the array curve in units on such a scale from array to array as to produce homoscedasticity because of the character of $F_1(y)$. That is, $F_1(y)$ and $F_2(x)$ are variable scales of measurement which when used to multiply standard deviations produce homoscedasticity. The method may be made clearer by dealing with some special cases.

(a) Given linear regression and constant homoscedasticity, we have $f_1(y) = m_1 y + c_1$, $f_2(x) = m_2 x + c_2$, where $F_1(y)$ and $F_2(x)$ are constants. The solution of the special functional equation leads to the normal correlation surface.

(b) Given linear regression and linear heteroscedasticity, we have $f_1(y) = m_1 y + c_1$, $f_2(x) = m_2 x + c_2$, and

$$\frac{1}{F_1(y)} = \lambda_1(y + a_1), \qquad \frac{1}{F_2(x)} = \lambda_2(x + a_2).$$

Then (8) takes the form

$$z = \Phi_1(y)\,\Psi_1\left\{\frac{1}{\lambda_1}\,\frac{x + g_1}{y + a_1} - \frac{m_1}{\lambda_1}\right\} = \Phi_2(x)\,\Psi_2\left\{\frac{1}{\lambda_2}\,\frac{y + g_2}{x + a_2} - \frac{m_2}{\lambda_2}\right\},$$

where $g_1 = m_1 a_1 - c_1$ and $g_2 = m_2 a_2 - c_2$.

* BIOMETRIKA, vol. 15 (1923), p. 224.

The solution of the special functional equation leads to

$$z = z_0(x+g_1)^{p_1}(y+g_2)^{p_2} \cdot \{(g_1-a_2)(y+a_1)+(g_2-a_1)(x+a_2)\}^q, \tag{9}$$

where p_1, p_2 and q are arbitrary constants. This is the general frequency surface with linear regression both ways and linear heteroscedasticity. The array curves both ways are Pearson curves of Type I. If $g_1-a_2 \to 0$ in (9), by simple transformations, the equation may be put into the form

$$z = z_0(x+g_1)^{p_1'}(y+g_2)^{p_2}e^{q'\frac{y+a_1}{x+g_1}}, \tag{10}$$

where $p_1' = p_1+q$, $q' = (g_1-a_2)q/(g_2-a_1)$. This special case has for arrays one way Pearson curves of Type III and the other way Pearson curves of Type VI.

(c) Given that the regression curves are certain equilateral hyperbolas both ways and that the standard deviations of arrays are of the form $1/(y+g_1)$ and $1/(x+f_2)$, respectively, it follows that the functional equation (8) takes the form

$$\begin{aligned} z &= \Phi_1(y)\Psi_1\{(y+g_1)x+f_1y+c_1\} \\ &= \Phi_2(x)\Psi_2\{(x+f_2)y+g_2x+c_2\}, \end{aligned}$$

and its solution is

$$z = z_0(x+f_1)^{\gamma_1}(y+g_2)^{\gamma_2}e^{\frac{\gamma(x+f_2)}{y+g_1}}, \tag{11}$$

which gives Pearson's Type III curves as array curves both ways.

The special cases we have given are to be regarded merely as illustrative. Narumi derives a considerable number of other surfaces. The publications of Narumi on this subject are to consist of Part I, II, and III. Part III has not thus far reached me.

13. *Extension of the Charlier System of Representation to Functions of Two Variables.* While great difficulties have been encountered in attempts to pass naturally from the Pearson system of generalized frequency curves to analogous surfaces for the characterization of frequency with respect to two variables, it appears that the way is

theoretically fairly clear for the extension of the Charlier system of representation to frequency functions of two or more variables. N. R. Jørgensen* has contributed to the solution of this problem. For simplicity, let us consider a continuous frequency function $F(x, y)$ of only two variables x and y measured from their respective mean values and in standard deviations as units as in § 6. Then the normal correlation function

$$\frac{1}{2\pi\sqrt{1-r^2}}e^{-\frac{1}{2}\Phi(x,y)},$$

where $\Phi(x, y) = (x^2+y^2-2rxy)/(1-r^2)$ plays a role in the representation of $F(x, y)$ analogous to that taken by the simple Gaussian function in the Charlier theory. The series to be considered is of the form

$$F(x, y) = \sum A_{m,n} e^{-\frac{1}{2}\Phi(x,y)} U_{m,n}, \tag{12}$$

where $U_{m,n}$ is a Hermite† polynomial defined by the equation

$$e^{-\frac{1}{2}\Phi(x-h,y-k)} = e^{-\frac{1}{2}\Phi(x,y)} \sum \frac{h^m k^n}{m!\,n!} U_{m,n}.$$

Let $\Psi(x, y) = (x^2+y^2+2rxy)/(1-r^2)$, and define $V_{m,n}$ by the equation

$$e^{hx+ky} = e^{\frac{\Psi(h,k)}{2(1-r^2)}} \sum \frac{h^m k^n}{m!\,n!} V_{m,n}.$$

In 1920, A. Guldberg‡ directed attention to the fact that the coefficient $A_{m,n}$ is readily expressible in moments of the given frequency function $F(x,y)$ by use of the well known fact that the functions $e^{-\frac{1}{2}\Phi(x,y)} U_{m,n}$ and $V_{m,n}$ form a biorthogonal system. That is, if we may assume the series (12) uniformly convergent, we find the coefficient

$$A_{m,n} = \frac{(1-r^2)^{m+n-\frac{1}{2}}}{4^{m+n}\pi m!\,n!} \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} F(x,y)\, V_{m,n}\, dx\, dy. \tag{13}$$

Since $V_{m,n}$ is a polynomial in x and y, the coefficients

* Loc. cit., p. 86.

† COMPTES RENDUS, vol. 58 (Jan., 1864); cf. *Oeuvres de Charles Hermite,* vol. 2, 1908, pp. 301–5.

‡ JOURNAL OF THE ROYAL STATISTICAL SOCIETY, vol. 83, p. 127.

$A_{m,n}$ are expressible as moments of $F(x,y)$. The series represents the function subject to conditions of continuity and vanishing at infinity given by Wera Myller Lebedeff.* The question of representing data by the first few terms of the series has not been answered and can probably be answered only by experiments with a wide range of distributions.

IV. The Theory of Random Sampling

14. *Introduction.* The theory of random sampling deals with problems of drawing inferences concerning the constitution of a statistical aggregate or class of things from the nature of a representative part taken as a random sample. The drawing of such statistical inferences about a class of individuals from the analysis of a sample is fundamental in the applications of mathematical statistics in insurance, in biology, and in other branches of science. That is to say, past experience with limited samples has been applied very widely to predict future events among the class from which the sample was drawn. But when we undertake the exact formulation of the theory which supports the applications even in the case of the simplest statistical ratios, we may find ourselves involved in the disputes about the validity of the theory of inverse probabilities. It is not my purpose here to enter upon a discussion of this well known controversy. I mention it simply because it seems desirable to direct attention to two recent interesting papers on the subject by E. T. Whittaker† and Karl Pearson‡ both published in 1920 as well as to the renewed attack on the theory by J. M. Keynes in his book published in 1921.

The Pearson school of statisticians has accepted the theory of inductive probability, and the statisticians of this school have been very active in dealing with the problems of sampling errors in various kinds of averages, statistical coefficients and parameters of frequency curves. Among

* Loc. cit., p. 415.

† Transactions of the Faculty of Actuaries, vol. 8 (1920), p. 163.

‡ Biometrika, vol. 13 (1920-21), p. 1.

the various sampling problems dealt with recently, the most interesting to me are the problems of the distribution of certain averages and coefficients obtained from small samples, and the extensions of the theorem of Tchebychef. I shall give the remaining time to these two topics.

15. *Fluctuations of Certain Averages obtained from Small Samples.* In the development of the theory of sampling, the assumption has usually been made that the sample contains a large number of individuals. But the lower bound of large numbers has remained poorly defined in this connection. For example, the usual probable error formulas have been applied to as few as ten observations. If there is a misapplication of formulas due to smallness of the sample, the source of the error would probably lie in the fact that the statistical constants from small samples are not distributed even roughly in accord with a Gaussian probability curve.

Beginning with a paper by Student* in 1908 there have been important experimental and theoretical results obtained on the distribution of arithmetic means, standard deviations, and correlation coefficients obtained from small samples. The material used in the experiments to which I refer consisted of 3000 pairs of measurements. The measurements were written on 3000 cardboards which were shuffled and from which 750 sets of 4 were taken. These provided two sets of 750 standard deviations each calculated from only four values, and 750 correlation coefficients each calculated from only four pairs of values. The distribution of the given 3000 values was roughly Gaussian in character. The simple inspection of the frequency distribution made it fairly obvious that the standard deviations experimentally obtained from sets of four were not distributed in accord with the Gaussian curve. Student found by empirical methods that the curve

$$y = y_0 x^{n-2} e^{-\frac{nx^2}{2\sigma^2}} \tag{1}$$

seemed appropriate to give the distribution of standard

* BIOMETRIKA, vol. 5 (1908), p. 1.

deviations s obtained from samples of n, where σ is the standard deviation of the infinite population from which the samples are drawn.

In 1915, Karl Pearson* took an important step in advance by obtaining the distribution of the standard deviations of samples of n variates from an infinite population distributed in accord with the Gaussian curve. He obtained from theoretical considerations the distribution of s identical with that which Student found experimentally.

Moreover, by tabulating β_1, β_2 and the measure of skewness for integral values of n from 4 to 100, Pearson shows that (1) approaches the Gaussian curve as n increases provided we accept certain necessary conditions, that is $\beta_1 = 0$, $\beta_2 = 3$, and skewness equal to zero, as sufficient for practical approach to this curve.

From this table, Pearson concludes that for samples of 50 the usual theory of probable error of the standard deviation holds satisfactorily, and that to apply it to samples of 25 would not lead to any error of importance in the majority of statistical problems. On the other hand, if a small sample, $n < 20$ say, of a population be taken, the value of the standard deviation found from it will be usually less than the standard deviation of the population.

Turning next to the Student† experiment with the distribution of the 750 correlation coefficients each computed from 4 pairs of values mentioned above. The correlation coefficient of the whole 3000 pairs from which drawings were made was $r = .66$. He further selected samples of 8 and samples of 30 from the population of correlation $r = .66$. An examination of the experimental results makes it fairly obvious that the distributions for samples of 4 and 8 pairs are far from normal, and that the average value of r from these small samples is smaller than the $r = .66$ of the total population of 3000 from which the small samples are

* BIOMETRIKA, vol. 10 (1915), p. 522.

† BIOMETRIKA, vol. 6 (1908-9), p. 302.

drawn. But with samples of 30, the correlation coefficient r approaches the correlation coefficient of the total population.

In a paper published in 1913, H. E. Soper* obtained to a second approximation the mean and standard deviation of the distribution of the correlation coefficient r from samples of n from a population of correlation ρ. He concludes that the mean value of the correlation coefficients obtained from small samples will be numerically less than the true correlation coefficient obtained from the aggregate and will be approximately represented by the formula

$$\rho\left(1-\frac{1-\rho^2}{2n}\right),$$

where n is the number in the sample.

In a paper published in 1915, R. A. Fisher† dealt with the frequency distribution of the correlation coefficient derived from samples of n pairs each taken at random from an infinite population distributed in accord with the normal correlation surface

$$z=\frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}}e^{-\frac{1}{2(1-\rho^2)}\left\{\frac{(x-m_1)^2}{\sigma_x^2}+\frac{(y-m_2)^2}{\sigma_y^2}-\frac{2\beta(x-m_1)(y-m_2)}{\sigma_x\sigma_y}\right\}},$$

where ρ is the correlation coefficient. The frequency function obtained for the distribution of r is given by

$$(2)\quad y_n=f_n(r)=\frac{(1-\rho^2)^{\frac{n-1}{2}}}{\pi(n-3)!}(1-r^2)^{\frac{n-4}{2}}\frac{d^{n-2}}{d(r\rho)^{n-2}}\left[\frac{\arccos(\rho r)}{\sqrt{1-\rho^2r^2}}\right].$$

The derivation of this form of $f_n(r)$ would probably be found of special interest to those who are seeking applications of certain general conceptions derived from the geometry of n-dimensional space. The ordinates y_n cannot be readily calculated from (2) except for small values of n. Moreover (2) offers no rapid means of calculating the mean value r, the modal value of r, or the standard deviation σ_r to replace the approximations obtained by Soper. In order to investigate

* BIOMETRIKA, vol. 9 (1913), p. 91.

† BIOMETRIKA, vol. 10 (1915), p. 507.

the approach of (2) to a normal curve as n increases, it seems necessary to provide methods for computing the ordinates y_n and moment coefficients for (2). This is accomplished in a joint memoir* by H. E. Soper, A. W. Young, B. M. Cave, A. Lee, and Karl Pearson. This memoir involves a tremendous amount of laborious numerical computation as well as the results of considerable theoretical work in adapting the function $y_n = f_n(r)$ and its moment coefficients to numerical calculation. The theoretical part consisted largely in obtaining series which converge with sufficient rapidity to be used in the numerical calculations. The tabulated values show the ordinates y_n at intervals of .05 for r from -1 to $+1$, at intervals of .1 for ρ from 0 to .9, and for $n = 3, 4, 5, \ldots, 25, 50, 100$ and 400, making in all 260 frequency distributions. The values of β_1 and β_2 were computed for these distributions to study the approach to the normal curve.

With respect to the approach of these distributions to the Gaussian form with increasing values of n, it is found that the necessary conditions $\beta_1 = 0$, $\beta_2 = 3$ for a Gaussian distribution are not well fulfilled for samples of 25 or even 50 whatever the value of ρ. For samples of 100, the approach to the conditions $\beta_1 = 0$, $\beta_2 = 3$ is fair for low values of ρ, but for large values of ρ, say $\rho > .5$, there is considerable deviation of β_1 from 0, and of β_2 from 3. For samples of 400, on the whole, the approach to the necessary conditions $\beta_1 = 0$, $\beta_2 = 3$ is close, but there is quite a sensible deviation from normality when $\rho \geqq .8$. These results give us a striking warning of the dangers of applying the ordinary formula for the probable error of r when we have small samples.

In conclusion, it should not be forgotten that the assumption is made, in this theory of the distribution of r from small samples, that we have drawn samples from an infinite population well described by a normal correlation surface, so that the conclusions are not in the strictest sense applicable to distributions not normally distributed.

---

* BIOMETRIKA, vol. 11 (1915-17), p. 328.

16. *The Tchebychef Theorem and its Recent Generalizations.* As the concluding topic of this paper, let us consider the recent extensions of the following remarkable theorem of Tchebychef which appeared* in the LIOUVILLE JOURNAL in 1867:

THEOREM. *If $a, b, c, \cdots$ represent the mathematical expectations of quantities $x, y, z, \cdots$ and $a_1, b_1, c_1, \cdots$ the mathematical expectations of their squares $x^2, y^2, z^2, \cdots$ the probability that the sum $x+y+z+\cdots$ is between*

$$a+b+c+\cdots+\lambda\sqrt{a_1^2+b_1^2+c_1^2+\cdots-a^2-b^2-c^2\cdots}$$

and

$$a+b+c+\cdots-\lambda\sqrt{a_1^2+b_1^2+c_1^2+\cdots-a^2-b^2-c^2\cdots}$$

will always be greater than $1-1/\lambda^2$, whatever the value of λ.

Tchebychef proved this theorem by simple algebraic methods. One great merit of the theorem lies in its freedom from restrictions with respect to the nature of the distribution of the variables.

To state the theorem in another form, let us assume the frequency distribution of an infinite population with standard deviation σ. If $P(\lambda\sigma)$ is the probability that a datum drawn at random from this distribution will differ in absolute value from the mean of the whole distribution by as much as $\lambda\sigma$, then

$$(3) \qquad P(\lambda\sigma) \leqq \frac{1}{\lambda^2}, \quad \text{or} \quad 1-P(\lambda\sigma) \geqq 1-\frac{1}{\lambda^2}.$$

In 1919, Karl Pearson† published an important generalization of the theorem subject to the mathematical condition that the frequency function $F(x)$ is such that the integral

$$\int_a^b (x-\bar{x})^{2s} F(x)\,dx$$

exists, where a and b are the lower and upper bounds of the distribution. He found

* *Des valeurs moyennes*, translated from the Russian by N. M. de Khanikof, JOURNAL DE MATHÉMATIQUES, (2), vol. 12, pp. 177-84.

† BIOMETRIKA, vol. 12 (1919), pp. 284-96.

$$1-P(\lambda\sigma) > 1-\frac{1}{\lambda^{2s}}\frac{\mu_{2s}}{\mu_2^s} = 1-\frac{\beta_{2s-2}}{\lambda^{2s}}$$

or

$$P(\lambda\sigma) \leqq \frac{\beta_{2s-2}}{\lambda^{2s}}, \tag{4}$$

where μ_{2s} is the $2s$th moment coefficient about the arithmetic mean of the area under $F(x)$. If we make $s = 1$, we have the special case of the Tchebychef theorem. In his proof, Pearson deals with moments μ_n where n is even. Seimatsu Narumi* has pointed out that it is sufficient to assume n a positive constant provided the integral

$$\int_0^\infty x^n\{F(x)+F(-x)\}\,dx$$

exists.

It is Pearson's view that, although his inequality is in most cases a closer inequality than (3), it is usually not close enough to be of practical assistance in drawing conclusions. Hence, it becomes important to obtain closer inequalities by decreasing the right hand side of (4). This was accomplished in papers published almost simultaneously by Birger Meidel† and B. H. Camp‡ by placing certain restrictions on the nature of the distribution function $F(x)$. But the restrictions are so devised as to leave the distribution function sufficiently free to be useful in the actual problems of statistics. The main restriction placed on $F(x)$ by Camp is that it is to be a monotonic decreasing function of $|x|$ when $|x| \geqq c\sigma$, $c \geqq 0$. The severity of this restriction on $F(x)$ varies according to the value of c. Its general effect is to exclude distributions which are not represented by decreasing functions of $|x|$ at points a certain assigned distance from the origin.

With the origin so chosen that zero is at the mean, Camp reaches the generalized inequality

---

* BIOMETRIKA, vol. 15 (1923), p. 246.

† COMPTES RENDUS, vol. 175 (1922), p. 806; cf. SKANDINAVISK ACTUARIETIDSKRIFT, 1922, p. 210–16.

‡ This BULLETIN, vol. 28 (1922), p. 427.

$$P(\lambda\sigma) \leqq \frac{\beta_{2s-2}}{\lambda^{2s}} \frac{\left(\frac{2s}{2s+1}\right)^{2s}}{1+\varphi} + \frac{\varphi}{1+\varphi} P(c\sigma), \tag{5}$$

where

$$\varphi = \frac{\left(\frac{c}{\lambda}\frac{2s}{2s+1}\right)^{2s}}{(2r+1)\left(\frac{\lambda}{c}-1\right)}.$$

With $c = 0$, the formula (5) is exactly Pearson's divided by $[1+1/(2s)]^{2s}$. With the origin chosen at the mode instead of the mean, and with moments defined with respect to the origin, Meidel* shows that

$$P(\lambda\sigma) \leqq \frac{1}{\left(1+\frac{1}{n}\right)^{n}\lambda^{n}}, \tag{6}$$

for any positive value of $n \geqq 1$. The general effect of the work of Camp and Meidel has been to decrease the larger number of the Tchebychef inequality by roughly fifty per cent. Even with the generalizations, the theorem of Tchebychef does not set such close limits on probabilities as the Gaussian law, but we should have regard for the fact that this theorem in its original form applies to any type of distribution, and in its generalized form to very general types of distribution.

In conclusion, it may be added that Markhoff† and Tschuprow‡ have in recent years given extensions of the work of Tchebychef on mathematical expectation and limiting values of probabilities along very different lines from those on which I have reported.

THE UNIVERSITY OF IOWA.

* Loc. cit., p. 214.

† A. A. Markhoff, *Wahrscheinlichkeitsrechnung*, 1912, p. 67.

‡ A. A. Tschuprow, *Zur Theorie der Stabilität statistischer Reihen*, SKANDINAVISK AKTUARIETIDSKRIFT, 1919.