PROCEEDINGS
OF THE
SECOND INTERNATIONAL CONFERENCE
OF
AGRICULTURAL ECONOMISTS

HELD AT
CORNELL UNIVERSITY,
ITHACA; NEW YORK,
AUGUST 18 TO AUGUST 29, 1930

The Collegiate Press
GEORGE BANTA PUBLISHING COMPANY
MENASHA, WISCONSIN
1930

3 cards

Research methodology

Agricultural economics Research

THEORY OF PROBABILITY AND ECONOMIC RESEARCH*

OSKAR N. ANDERSON
HANDELSHOCHSCHULE, VARNA, BULGARIA

ACCORDING to a report in the June number of the Journal of the American Statistical Association, Professor Warren M. Persons made the following statement at a dinner meeting of the American Statistical Association, held at the Aldine Club, New York City, February 13, 1930:

> "Statisticians may be divided into two broad classes. The first class, consisting of men like Oscar Anderson and Arne Fisher, are fundamentalists. The fundamentalists believe in a trinity: mathematics, probability, correlation. They are more interested in nice mathematical processes than in an examination of premises. They believe in absolutes. Their point of view is mechanistic. The second class, consisting of such men as Bowley, Keynes, Allyn Young, and E. B. Wilson, are skeptics. The skeptics include among their number many of the greatest mathematical statisticians. They conceive mathematical statistics to be similar to mathematical physics in which, to obtain realistic results, premises must be subjected to the closest scrutiny. They value mathematics as an instrument, but they are not hypnotized by mathematical symbols and they do not stand in awe of the tool they use. They do not believe in iron clad laws or mechanistic interpretations in economics. They suspect, for instance, the validity of correlation coefficients derived from economic data by involved processes."[1]

I must confess I am not a little surprised to find myself placed by Professor Persons in the same row with the "fundamentalists"—if they really exist. I have examined the premises of the statistical methods used by Professor Persons, and have called attention in my writings to what I believe to be weaknesses in certain of these premises. Furthermore, it is possible that Dr. Persons himself might be accused of being a fundamentalist with his trinity: the method of least squares, link relatives, and the elimination of the trend.

* This paper was translated from the original German manuscript by Oswalt Vopelius, and O. B. Noltenius of Cornell University. The translation was subsequently checked by Professor Anderson, and returned with the request that someone read the paper and make such further corrections as appeared to be necessary. This was done by S. J. Pretorius and F. F. Hill of Cornell University . Unfortunately, time did not permit of sending the corrected manuscript to Professor Anderson for a second reading. It is sincerely hoped that no serious errors have been introduced.

[1] Journal of the American Statistical Association, Vol. XXV, New Series, No. 170, June, 1930, page 208.

However that may be, Professor Persons' remarks as reported in the Journal of the American Statistical Association, lead me to explain in greater detail the viewpoint of the so-called "Stochastic School" with reference to the question of the relation of the theory of probability to economic research. As this question is also of importance in the field of theoretical research in agricultural economics, I think it is entitled to some of the time of this Conference.

As is generally recognized, economics is the only science in which there is open conflict between the adherents of the mathematical and of the non-mathematical points of view, and in which those who are not blessed with mathematical ability, or at least those who have not had the opportunity to learn mathematics, put their ignorance on a pedestal and publicly boast of it. Instead of choosing a field of study in which a knowledge of mathematics is unnecessary, they deny the application of mathematical methods in the field of economic research, even though admitting they know nothing of such methods, and declare all writings on economic subjects which contain mathematical symbols to be not worth reading. While the above statements do not apply to all European countries (Germany in particular must be excepted), they certainly apply in a large measure to America.

There should be no more conflict between mathematical and non-mathematical groups in the field of economic research than there is conflict between "microscopists" and "non-microscopists" in the field of natural science. Mathematics is only an instrument of research, a method of thinking, which is applied when other logical processes of thinking fail. If a "skeptic" is defined as a person who uses mathematics as a tool in economic research, then I am willing to be called a skeptic.

I should like to emphasize that in at least one respect I am more of a skeptic than Dr. Persons himself, for I do not believe that mathematical statistics, with its descriptive and analytical methods, can ever be substituted for economic theory. For example, the refined methods of mathematical statistics such as the decomposition of series into separate components, the calculation of coefficients of correlation, and even the computation of means and standard deviations, rely upon premises and hypotheses which still remain to be proved. I believe that in every instance it is essential to determine whether or not such statistical methods are applicable to economic phenomena. If such methods are employed without first determining whether or not they are strictly applicable to the particular problem under consideration it frequently happens that daring, if not even false assumptions are smuggled in, although the investigator may pretend to be truly statistically descriptive. You may for example eliminate the trend. But what is the trend? The answer is that it is something I have determined upon; some *X* in a time series which, diagrammatically, we express as a straight line or a parabola of a lower order. Please excuse my saying so, but this is a geometrical definition, rather than an economic one. Furthermore, a question may be raised as to whether excluding this *X* does not introduce certain audacious assumptions as to the behavior of various economic phenomena.

The relation of statistics to theoretical economics is, I believe, analogous to the relation of experimental physics to theoretical physics. The physicist builds up his hypothesis, either (1) on the basis of deductions, dependent upon his knowledge of the properties of the object under consideration, (2) on the basis of inductive reasoning as in the case of Newton's Law, and Ohm's Law, or (3) on the basis of systematic experiments in a determined direction. In every case, however, the hypothesis is carefully tested to determine if it is in agreement with all of the known facts.

In the field of social science, it is very seldom possible to experiment; in fact scientifically exact experiments are really out of the question. We content ourselves, therefore, with such substitutes as the statistical method provides: we determine ratios and averages, construct correlation tables, resolve series into their component parts, calculate coefficients of correlation, coefficients of variation, standard deviations, and so forth. However, as previously stated, the use of such methods does not of itself provide a substitute for theoretical economics. Statistical methods should only be used for verifying or checking theoretical conclusions. If, therefore, Dr. Persons does away with the distinction between theoretical economics and statistics, and believes that "the various theories of business fluctuation are of little use in the practical problem of business forcasting," I must say that I cannot agree with him.

I do not wish to imply, of course, that I regard all theories as providing equally fertile hypotheses. In this connection Altschul aptly points out that "In such an opinion (namely, that all economic theories do provide equally fertile hypotheses) one can, and not unjustly, find confirmation of the view frequently held in Germany that the empirical statistical school of American economists would mean, fundamentally, a revival of the school of Schmoller, with the only difference that statistical analysis in the guise of mathematics would have taken the place of historical description."[2]

The irony of Professor Persons is probably directed toward those mathematical statisticians who wish to apply the theory of probability in economic research, or, to be more exact, who wish to verify economic theory by means of the theory of probability. There is no doubt that at the present time there is a sharp difference of opinion between the school of thought which Dr. Person represents, and the "stochastic" school of thought. The latter group, as for example the followers of the late Professor Tschuprow of Russia, is now beginning to center around the Frankfurter Society for Konjunktur Research. The differences of opinion between the two schools is, to a certain extent, based upon misunderstanding. I intend to point out in this paper that the use of the theory of probability in economic research cannot be regarded merely as a "caprice" of those who "stand in awe of mathematical symbols," but that its use is a matter of logical necessity. In the following discussion I shall have to present certain of the material from my recently published book "Die Korrelationsrechnung in der Konjunkturforschung."[3]

The statistical data which we are forced to use in practice are seldom absolutely accurate. Such data, for the most part, contain errors of varying degrees of magnitude. The true figure in which we are interested, as for example, the population of Bulgaria at 12:00 P.M., December 31, 1926 (which in this case we shall designate as A) is seldom available. Instead, the Census reports a somewhat different figure which we shall designate as A'. If we assume the difference between A' and A, $(A' - A)$, to be represented by e, then e is the error of our determination, and we

[2] Altschul, E., Die Mathematik in der Wirtschaftsdynamik, grundsätzliche Bemerkungen; Archiv für Sozialwissenschaft und Sozialpolitik, Band 63, Heft 3, 1930, page 524.

[3] Oskar Anderson, "Die Korrelationsrechnung in der Konjunkturforschung. Ein Beitrag zur Analyse von Zeitreihen." Bonn, 1929. (Veröffentlichungen der Frankfurter Gesellschaft für Konjunkturforschung, herausgegeben von Dr. E. Altschul, Heft 4)

may write, $A' = A + e$. These "errors of observation" occur in all sciences in which observations are made. In order to cope with the problems arising out of errors of observation, a special science known as the Theory of Errors has been built up. As is well known, it is based upon the theory of probability.

The theory of errors, at least up to the present time, has been developed primarily to meet the needs of astronomers and physicists. The physicists are usually in the fortunate position of being able to set up their experiments in such a way that the errors represented by e are relatively small as compared with A, and can therefore be eliminated by the use of relatively simple means. Such small errors rarely occur, unfortunately, in the "observations" of social phenomena. If the statistics or "observations" are well organized and *complete,* as is more or less the case with population statistics and statistics of births and deaths, the errors are, of course, much smaller. Statistics of business and trade are much less accurate than population statistics, and it may be said that the majority of economic statistics (including those relating to agriculture) contain many sources of error. Unfortunately, this fact is absolutely ignored by most theoretical economists. The sources of error, for the most part, differ from those in the fields of physics or even biology, and the entire subject is much more complicated. The methods of the theory of errors, which, as stated above, are based upon the theory of probability, must be adjusted, modified, and added to, before they can be used. In most cases the single "elementary errors" which combine to make up the "total error" occur less frequently here than in the field of natural science, but their magnitude is greater, and their "laws of distribution" are frequently such that they admit only of the application of Tchebycheff's inequality. The development of a satisfactory application of the theory of probability to the problems of economic research, is one of the first problems which mathematical statisticians have to solve.

Let us assume that during k moments of time the prices of a certain good were registered in the market as follows:

$$A'_1,\ A'_2,\ A'_3\ \ldots\ldots\ldots A'_k$$

As previously pointed out, these prices can also be represented as follows:

$$A_1 + e_1,\ A_2 + e_2,\ A_3 + e_3, \ldots\ldots\ldots\ldots A_k + e_k.$$

If A_1, A_2, A_3,........A_k are equal to one and the same quantity A, the arithmetic mean results from the k actually observed quantities, A'_1, A'_2, A'_3,........A'_k as

$$A + \frac{e_1+e_2+e_3+ \ldots\ldots\ldots\ldots e_k}{k}$$

This value, which, under certain circumstances, results in a considerable error in the determination of A, cannot be used unconditionally until certain hypotheses relative to the size and distribution of the errors e_1, e_2, e_3,..........e_k are introduced—hypotheses which must be based upon the theory of probability. In the classical theory of errors it is usually assumed that the errors are independent of one another and that they are as likely to be positive as negative. If this assumption is correct the expression

$$\frac{e_1+e_2+e_3+ \ldots\ldots e_k}{k}$$

becomes small as compared with A, and the arithmetic mean of the k "observations" closely approaches the true value, A. But the foregoing assumption relative to the behavior of single errors of "observation" is only a hypothesis, and it cannot be assumed to be unquestionable in the field of economic phenomena, as some people apparently believe. There are not a few cases when positive and negative errors fail to balance, and when, as a matter of fact, the errors are actually correlated. Furthermore each A_1, A_2, A_3,........A_k, cannot be considered to be equal. This leads to the difficult problems of the so called "smoothing" theory. If other hypotheses are not introduced relating to the expected character of the errors of observation, hypotheses which have to be subjected to the closest scrutiny with respect to whether they are justifiable or not, the limits of error of the computed averages will be considerably greater. The more such averages are combined and recombined in making a further analysis, the greater are the errors carried along through the computation, and the more distorted are the results. Such errors can easily occur in making determinations of demand curves.

It is characteristic of the mathematicians of the Lausanne school of political economy, probably blinded by their differential equations, that they usually do not apply the theory of probability when they want to find certain approximations to their theorems. The justification for such approximations is based on the actual material; however, the material may include large errors.

The theory of probability might also be applied with advantage in the field of economic research to the problem of sampling. There was a time when Georg v. Mayr, for instance, defined statistics as "complete" mass observation. However, developments during the past four decades have demanded so much from the science of statistics, and the whole field has become enlarged and complicated to such a degree, that one is obliged, at least in so far as economic statistics are concerned, to depend upon those incomplete "substitutes" for the statistical observations which a Laplace once imagined, and which have been used, in part at least, by the political "arithmeticians."

One of the most important of these "substitutes" is the process of sampling. As is well known, sampling may be done in either one of two ways; (1) in the conscious choice of a certain number of typical representatives from a statistical universe which is to be described or (2) in the "random" choice of such representatives. The method of random sampling is entirely dependent upon the applications of the theory of probability, and the great advantage of this method lies in the fact that one can compute in advance the unit within which the errors of observation lie. It may also be said that Bowley and Jensen have introduced the theory of probability in the process of sampling by the first method mentioned above.

In support of the practice of using the sampling process in making statistical analyses, it may be argued that:

1. In certain instances it is, in the very nature of the case, impossible to obtain data covering all of a large number of cases It is impossible for all practical purposes, for example, to secure a complete census of *all* family budgets.

2. There are many cases where absolute accuracy is not essential, and where all that is necessary is to make certain that the errors of observation do not exceed a certain limit. A chemist's balance is certainly much more accurate than a common scale, but a load of hay purchased on the market, is never weighed on the former.

3. It is very seldom that the results of even so-called complete

statistical observations are exact. They contain, as has been previously pointed out, errors which may run as high as 20 per cent. A census of real estate in countries which do not have a regular land register, as for instance in Bulgaria and Russia, is a case in point.

4. Even if a census is absolutely accurate as of the day it was taken, it cannot be considered as entirely correct on the day of its publication, since changes are certain to have occurred in the meantime. In order to estimate the amount of such changes we have to use the rather unsatisfactory method of *extrapolation.*

It is probably safe to say that, at the present time, the major part of the material used in economic statistics (including agricultural statistics) is incomplete, and at its best made up of "representative" observations. Crop statistics and crop forecasts in which the actual or even estimated yields on *all* farms are never obtained are cases in point. Index numbers of prices are never based on all prices of all goods in all localities.[4] Investigations of farm incomes, economic enquêtes, and unemployment statistics are likewise never based upon complete data. It is probable that in the future the proportion of all investigations in the field of economics, which will be based upon more or less complete data, will be even smaller than it is today.

Results based upon samples can but be considered as approximations to the characteristics of the complete body of data, the actual characteristics of which remain unknown to us. The foregoing statement is especially true of price indices if they have any economic significance at all. In general the problem is more or less as follows: A certain economic field is given in which during a certain period in the past which we shall designate as T, N different items of goods were sold at different prices. We are interested in a certain function F of these prices and amounts of goods which we shall call the true index number. From all of the N items of goods we select a certain number n and taking these, together with their prices (which prices have really been

[4] I cannot understand how the theory of index numbers could exist without the theory of sampling. However there are theorists enough who believe there is no use for it. One of the new editions of the "Zeitschrift für Nationalökonomie", Vienna, will contain an article which I have written on the question, "Is it possible to prove the quantity theory by the use of statistical methods?" In this same article I also point out the sampling character of the wholesale index in the well known equation of Irving Fisher.

paid during the period T for the items chosen), we construct a function f and consider it as an empirical approach to the function F, which, as previously stated, remains unknown to us. It is quite evident that in this way our problem can be reduced so that the conditions are quite similar to those involved in the classic experiment of sampling ballots from boxes. It makes no great difference whether or not we select our group of items at random. It must be recognized that in general in applying the method of random sampling, one can simply resort to the corresponding box schemata of the theory of probability. In brief, if we are dealing with results based upon data selected by a process of sampling, we must always take into consideration the limits of the errors of sampling, and in doing this we are in the field of the theory of probability. Failing to recognize this, we are, consciously or unconsciously, playing hide and seek with the theory of probability, which, of course, does not do much toward advancing our science.

Use is made or should be made of the theory of probability in making forecasts, such as crop forecasts, price forecasts and so forth. It is my opinion that such forecasts should be expressed as follows: Forecasted price, for example, $1.00 per bushel; possible margin of error ± 60 cents (sometimes more, sometimes less). The margin of error depends of course on the accuracy of the determination and on the number of elements taken into account in making the forecast. It is recognized more and more that the theory of probability must be applied in making such forecasts. It is a real pleasure, in this respect, to turn over the leaves of the more recent issues of the Journal of the American Statistical Association, and to note the shaded areas which represent the probable l'mits of error of the extrapolated curve.[5] I believe that if similar methods were used in forecasting business cycles, many disappointments would be avoided.

Many more cases could be cited where the application of the theory of probability would seem to be necessary, but the examples already given make clear that in most economic research where statistical data are being used to verify a theoretical hypothesis, the theory of probability should be applied. Most statistical data

[5] See for example, "The Standard Error of Forecast from a Curve," Henry Schultz, Journal of the American Statistical Association, Vol. XXV, New Series, 170, June, 1930, page 139.

contain errors and can be considered only as empirical approaches or approximations to the true figures in which we are interested, which figures are not available. The true figures in which we are interested may, from the standpoint of the theory of probability, be called *a priori* values.

However, it would be a great mistake, unfortunately often committed, simply to transfer to the field of economic research the methods of applying the theory of probability which have been worked out in other fields; especially in the field of biometrical mathematical statistics. In time series, where successive observations may be closely related, the situation is entirely different from that usually found in the field of biometry. The situation can be successfully dealt with only if the assumptions upon which biometrical methods are based are carefully examined and only if, in certain instances, those methods are adjusted to meet the new situation, or new methods developed to meet it.

The stochastic school believes that it would be of considerable advantage to start from a system of conceptions of theoretical probability most of which has been developed by a few prominent Russian mathematicians and statisticians. Under this system the "algorithm" of the "method of mathematical expectations" can be usefully applied.[6] Here we have the conception of the *random variable.* Everything statistically comprehensible—amount, total value, average, ratio and so forth—is expressed in terms of values which change, or at least can change, with time. For example, the population of a country increases, the death rate decreases, prices rise and fall and so forth. The statistical figure related to a certain object represents, if considered in time, a variable figure or briefly a *variable* as we usually call such figures in mathematics. As has been pointed out several times, such statistical figures are seldom perfectly exact. They are, for the most part, influenced by major or minor errors. To cite

[6] The Russian School has always been very exact in their definitions and in establishing premises which are required for making theoretical deductions. In my own work, I have tried, as far as possible, to continue this practice, although it has, upon occasions, left me open to attack by my critics (Lorenz, Tinbergen and others). It is much easier to attack hypotheses and premises which are openly presented than to point out errors in the premises and hypotheses of applied methods which have not been carefully and systematically put forward. One sees the thorn in the eye of one's neighbor but overlooks the wood in one's own eye. Nevertheless I do not envy the laurels which can be won—for a short time, at least—by working out new statistical methods, the theoretical bases of which have not been carefully examined.

an example already used, the population of Bulgaria as of January 1, 1927, as reported by the Census, was 5,483,125. It can certainly be assumed that the actual population within the boundaries of Bulgaria at 12:00 P.M., December 31, 1926, was somewhat different from the figure reported. Just what the difference between the *true* figure and the figure reported amounted to we do not know. We can only guess at the difference, which means that we must establish a hypothesis. We can, for example, assume that a negative error was more probable than a positive error, that is to say, that the population of Bulgaria was somewhat greater than 5,483,125 at 12:00 P.M., December 31, 1926. We may assume further that the error was not large or that a large error was less probable than a small one, and so on. In other words, it would be possible to write down a series of figures, one of which would certainly be the correct population figure at 12:00 P.M., December 31, 1926. This figure might be 5,483,125, 5,483,126, 5,483,127, *et cetera.* We could assign to each of these figures a certain mathematical probability that it represents the true population figure as of the above date, as for example 1/100, 1/150, . . . 1/1,000, and so forth. If we were to write down all the population figures that are possible according to the given conditions, together with the probabilities assigned them, then, if we had estimated correctly, the sum of the probabilities would be equal to unity, since one and only one of the estimates of the population is certain to be correct. By this process, however, we have assigned to the population figure the property not only of a variable but of a random variable for "a quantity which have *K* different values with definite probabilities is called a random *variable of the order K*" (Tschuprow). All of the possible values together with their corresponding probabilities constitute the so-called "distribution law of the variable." In the above sense, almost all statistical quantities may be considered as random variables, for most of them are subject to errors of observation. Certain ranges may be assumed, or at least thought of, within which their true values must lie, and a certain probable error may be thought of for each of the possible values within the range. In general, such probabilities are only *thought* of as existing. The concept of a random variable is equally valid whether the variable actually exists or not. The concept of the "random variable" is logically and mathematically refined, and, of course, represents an abstraction or transformation of the observed facts, just as a geometrical figure represents an abstraction of actually existing figures. The same applies to nearly every scientific concept.

We can distinguish different kinds of random variables. There are random variables which result from errors of observation in measuring a constant. Again there are variables, where the measured value itself is a random variable. Furthermore, there are variables which correspond to something which actually exists, and other variables which may be regarded as statistical abstractions, such as averages, ratios, and so forth.[7]

Certain of the random variables which appear as statistical ratios, have the characteristics of empirical frequencies, as for example, the percentage of the farms in Bulgaria with an area of 0-1 hectares. This percentage amounted to 11.78 per cent, or 1,178/10,000 of the 80,565 farm census cards which were selected from the 734,769 cards correctly filled out in the agricultural census of December 31, 1926—January 1, 1927, and was "representatively" worked out under my supervision.

The concept of empirical frequencies may be illustrated by the use of a closed urn containing a certain number of ballots or balls exactly alike in all respects except as to color. Let us take, for example, an urn containing 100 white and 200 black balls which are exactly alike in all other respects. If the balls are drawn from the urn one at a time, the mathematical probability of obtaining a white ball in the first drawing is, of course, expressed by the ratio $100/(100 + 200) = 1/3$. If in 100 drawings we get 35 white balls and 65 black balls, then $35/100$ or $7/20$ is the empirical relative frequency with which white balls were drawn. We know further that according to the "law of large numbers" the empirical relative frequency will approach the mathematical probability (in our case the fraction $1/3$) the greater the number of drawings which are made. Thus the mathematical probability may be regarded as a limit to the empirical relative frequency (Mises). In the example used above, the limit of the percentage of small farms in Bulgaria was 11.76, which means that this represents the

[7] See my "Korrelationsrechnung," pages 15-17. I cannot agree, however, with the statement of Lorenz (Jahrbüchern für Nationalökonomie und Statistik, III, F., 78 Band, 1 Heft, July, 1930, pages 148-151) that a random variable, in the real meaning of the word, is the same as a random variable taken as meaning the result of chance. With such statements one could totally destroy in a moment the whole of the theory of errors when once the proof has been worked out.

percentage of the total 734,769 farms reported to the census, which contained one or less than one hectare. The difference, *e*, between this value (11.76 per cent) and the "representative approximation" (11.78 per cent) is, therefore, only 0.02 per cent or 2/10,000.

The limit under consideration must not be confused with the limit of a mathematical series, for the latter has nothing whatever to do with the idea of probability. The approach to the limit in our case is not regular, but rather it occurs through irregular zigzag jumps. The empirical relative frequency is itself a random variable. It may amount to exactly 1/3 after only three drawings, in which case greater accuracy could not be secured after a billion drawings (a billion is not divisible by 3, and an empirical relative frequency with a billion in the denominator can therefore never reduce to precisely 1/3). Consequently a special term has to be introduced for the limit of a random variable. This limit value differs, as previously noted, from the usual use of the term through being associated with the theory of probability. Following the proposal of Slutsky we shall call it the "stochastic limit" and designate it by the symbol $\lim_B$ (Bernoullianus limes). A complication, which is important from the standpoint of practice, arises if the stochastic limit, to which our empirically given quantity tends, is itself a variable which changes directly with the number of observations. In this case we do not use the term stochastic limit but the term "stochastic asymptote" (as $_B$= Bernoulliana asymptota).[8]

In statistical practice we have to deal much more frequently with arithmetic averages than with empirical frequencies. They also have a stochastic limit (limes or asymptotes). This limit appears as their mathematical expectation. The concept of mathematical expectation is, as has been mentioned above, the basis of an entire method and is very often applied by Russian mathematical statisticians. In spite of the German works of Bortkiewicz and Tschuprow, the use of the term in the above sense is not generally known to English and German statisticians. The French, however, seem to be more familiar with it (see G. Darmois: Statistique Mathématique, Paris, 1928). Furthermore, as its use in the field of economic statistics has brought forth criticism from various quarters it is necessary to discuss the subject in some detail in order to make its meaning entirely clear.[9]

[8] Anderson, Oskar, "Korrelationsrechnung," pages 18-20.

[9] See the article of Dr. P. Lorenz, "Der Begriff der Mathematischen Erwartung

A mathematical expectation of a random variable (if we remember that in economic statistics most figures may be considered as random variables within the meaning of the theory of errors) denotes the sum of the products formed by multiplying all values of the variable by their respective probabilities, or in other words, the weighted arithmetic mean of all values of the variable when each value of the variable is weighted by its probability (the sum of these probabilities is always equal to 1. Therefore the denominator in the expression of the weighted average disappears.) The mathematical expectation is usually designated by the symbol *E*. The mathematical expectation of the square of *x* is therefore expressed as $E\,x^2$.

The concept of mathematical expectation may be looked at from another point of view. Let *X* be a random variable which can have only *k* different values, $x_1, x_2, x_3, \ldots . x_k$. Suppose that *N* experiments have been made and that the value of x_1 has been observed n_1 times, that the value x_2 has been observed n_2 times, the value x_3, n_3 times, and so on. Consequently, $n_1 + n_2 + n_3 + \ldots\ldots = N$ The usual weighted arithmetic average of the *N* experiments is equal to

$$\frac{I}{N}(n_1x_1+n_2x_2+n_3x_3+\ldots\ldots\ldots\ldots) = \frac{n_1}{N}x_1+\frac{n_2}{N}x_2+\frac{n_3}{N}x_3+\ldots.$$

The quotients· $\frac{n_1}{N}, \frac{n_2}{N}, \frac{n_3}{N} \ldots\ldots\ldots\ldots$

are here the empirical relative frequencies of the single values of the variable *X*. If we replace these relative frequencies by the corresponding mathematical probabilities, $p_1, p_2, p_3, \ldots\ldots\ldots$, associated with the corresponding values of *X* we arrive at the

in Statistik und Konjunkturforschung" in den "Jahrbüchern für Nationalökonomie und Statistik," III F, 77 Band, 6 Heft, June, 1930, pages 832-843, and his above mentioned review of my book. Lorenz is collaborator at the "Institut für Konjunkturforschung" (Professor Wagemann) and appears to be a very good mathematician. But the works of the so-called Stochastic School, except my own two monographs are apparently unknown to him, especially the works of Tschuprow. It is interesting to note that his criticism might very properly be directed against the theories of Professor Wagemann himself, as the latter deals with "stochastic connections" and so on (see for example his "Einführung in die Konjunkturlehre," Leipzig, 1929, page 42 et. seq.). I hope that both Lorenz and Wagemann are unconscious of this disagreement.

following expression for the mathematical expectation of the variable X:

$Ex = p_1x_1 + p_2x_2 + p_3x_3 + \ldots\ldots\ldots + p_kx_k$. Let us illustrate this by a concrete example. Suppose we have to determine the average yield of corn per hectare in Bulgaria during the year 1930. This quantity, which we shall call A, can obviously be defined as the quotient of the total number of quintals of corn harvested divided by the total number of hectares in corn. Actually, the figures as to yield and the area in corn are unknown, and we are bound to compute a figure A' which is only an approximation to A. We do this by selecting a small number of fields in different parts of the country the yields of which were b_1, b_2, b_3 .... quintals per hectare. We multiply the yields by the number of hectares in the respective fields $(a'_1, a'_2, a'_3 \ldots\ldots\ldots)$ and divide the sum of the products by the total number of hectares in the fields selected. The mathematical formula is expressed as follows:

$$A' = \frac{\Sigma a'_i\, b_i}{\Sigma a'_i}$$

or if we write $\Sigma a'_i = a'$

$$\text{then, } A' = \frac{a'_1}{a'}\, b_1 + \frac{a'_2}{a'}\, b_2 + \frac{a'_3}{a'} b_3 + \ldots\ldots\ldots\ldots$$

The quotients $\frac{a'_1}{a'}, \frac{a'_2}{a'}, \frac{a'_3}{a'} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots,$ have the character of empirical relative frequencies. For instance $\frac{a_i'}{a'}$ represents the relative frequency of occurrence of a certain yield per hectare, b_i, among all yields per hectare, actually observed, or measured, in 1930. To simplify the problem we have assumed that all of the yields b_1, b_2, b_3 .... are different. If they are not all different, all yields per hectare occurring more than once, *i.e.*, all cases where the value of b is the same, may be grouped together, the corresponding area, of course, being represented by the sum of the areas of the individual fields.

It is possible according to the formula for A', to change also the formula for A. Assume that the actual yields per hectare are given by the series b_1, b_2, b_3, b_4......, where the yield b_1 was on a_1 hectares, the yield b_2 on a_2 hectares, and so on. Assuming further that, $\Sigma a_i = a$, then it follows at once, on the basis of the assumptions, that:

$$A = \frac{\Sigma\, a_i b_i}{\Sigma a_i} = \frac{a_1}{a} b_1 + \frac{a_2}{a} b_2 + \frac{a_3}{a} b_3 \dots\dots\dots\dots$$

It is fairly obvious that the expressions, $\frac{a_1}{a}, \frac{a_2}{a}, \frac{a_3}{a}, \dots\dots\dots\dots,$ play the part of mathematical probabilities with respect to the empirical relative frequencies, $\frac{a'_1}{a'}, \frac{a'_2}{a'}, \frac{a'_3}{a'}, \dots\dots\dots\dots\dots\dots\dots\dots\dots$

For instance $\frac{a_1}{a}$ represents the probability that one of the selected fields yielded b_1 quintals per hectare while the empirical relative frequency $\frac{a'_1}{a'}$ is only an empirical approximation to this quantity and may also be considered as its stochastic limit. Therefore, the required quantity A, according to the above definition, is equal to the mathematical expectation of A'. This is due to the fact that in the expression of the weighted arithmetic average of A', mathematical probabilities are substituted for all corresponding empirical relative frequencies.[10]

The conception of "mathematical expectation" is of considerable importance in all statistical theory and it should be applied as far as possible, in practice. It is apparently not well known to most statisticians, and its name, while historically justified is not well chosen. But similarly, the differential coefficient $\frac{dy}{dx}$ appears

[10] If a field of a certain size does not occur in our sample, though we have made provision for it in our formula, then its empirical frequency in A is obviously to be taken equal to zero.

as a somewhat artificial form to many beginners in the study of higher mathematics, and it is only during the course of their studies that its whole meaning becomes clear to them.[11]

As stated above, the arithmetic average is related to its mathematical expectation as to its *ideal* or *stochastic limit,* just as the empirical relative frequency is related to its corresponding mathematical probability. One can always proceed from mathematical expectation to mathematical probability by simply assuming that the happening of a given event could have but one of two values, namely, 0 or 1; its mathematical expectation will then be equal to its mathematical probability. Furthermore, from a certain point of v.ew every mathematical probability may be looked upon and treated as the mathematical expectation of the empirical relative frequency corresponding to it. The system of theorems on mathematical expectations comprises also *in nuce* the entire system of the corresponding theorems on mathematical probabilities. In the majority of cases in actual practice, only the mathematical expectation will occur as the stochastic limit of any random variable. The introduction of the conception of mathematical expectation is further favored by the fact that the technique required for its use is usually quite easy to handle. Some very general, and at the same time, very simple theorems of the mathematical expectation enable us to treat it, not as the sum of a large number of products of unknown quantities, but as a symbolic operator; something similar to the sign of the integral, etc. Quite generally, for example, the mathematical expectation of a sum of constant or variable quantities equals the sum of the mathematical expectations of the separate items; the mathematical expectation of the products of variables, "stochastically independent" of one another, equals the product of the mathematical expectations of the separate variables, and so on.[12] A calculation involving mathematical expectations requires therefore a special procedure, or as mathematicians sometimes say, a special algorithm.[13]

[11] Those who are acquainted with the work of the biometrical school of Karl Pearson, will readily recognize that in the example used above, the mathematical expectation is identical with the so-called weighted mean for the whole population. However, the mathematical expectation, and the weighted mean for the whole population need not always be the same.

[12] Two random variables are said to be "stochastically independent" of each other if, after fixing a value for one variable, the law of distribution of the other variable is not changed in any respect.

[13] Anderson, Oskar, Korrelationsrechnung, pages 20-21 and 101-103.

Lorenz, whose criticism on the whole is well grounded and to the point (in spite of the many exclamation marks) finds the conception of the mathematical expectation of a single variable, at least in the case of a time series, "rather nebulous."[14] However, I can not quite understand this objection. If A, the *true* average yield of a hectare of corn, equals the mathematical expectation of its empirical approximation A', it is entirely irrelevant whether this A' does or does not repeatedly appear in the series in question. If there is any sense at all in regarding a statistical figure X' as an empirical approximation of a figure X upon which it is based, but whose exact value we do not know, then $e = X' - X$ is the error of our calculation and we may write:

$$X' = X + e$$

On the supposition *rebus sic stantibus* (something like the known "statics" of the Lausannians), it may be assumed that the "error e" which is under the influence of different causes has a certain "law of distribution," that is, it might have different values with different probabilities—always under the assumption *rebus sic stantibus.* Not only the Gaussian law of errors, but also most of the other symmetric or moderately asymmetric laws of distribution have the property that the mathematical expectation of the error e equals 0, and therefore we have:

$$EX' = E(X + e) = EX + Ee = EX = X$$

because, as we assumed, X itself is a random variable. And even if Ee does not equal 0 but a constant magnitude c, the c will also disappear with the calculation of the standard deviation, the correlation coefficient, and similar coefficients where deviations from the arithmetic mean are taken.

It is true that the value of X cannot be obtained from a single observation X', but if a number of observations arranged in a time series are available where the mathematical expectations of each observation may differ, then on the basis of certain assumptions a series of approximations to the value of X which are better than

[14] I found in Lorenz' review only one assertion which is really wrong, namely the statement that in my second example "upon the relation of wheat prices in Berlin and in New York (Korrelationsrechnung, pages 91-96) there is an unintelligible, *and in the text, unexplained,* coefficient of correlation of –0.4" (see pages 150-151 of the review). A very detailed explanation of this coefficient is given in my book on pages 94-95.

the original values of X', may be obtained. This process is involved in the so-called smoothing of series.[15]

The conception of mathematical expectation may find another and greater application in the field of economic research. This application is the object of a great deal of criticism and probably will continue to be for some time—until it has been proved to be either correct or scientifically sterile.

From an urn containing M white and N black balls, m white and n black balls are drawn and put aside. From the $m + n$ balls thus obtained m' white and n' black ones are withdrawn. Here $\frac{m'}{m'+n'}$ is the empirical relative frequency of drawing a white ball in the second experiment and may be regarded as an empirical approximation to the probability $\frac{m}{m+n}$ underlying it. This probability, however, is only an empirical relative frequency with respect to the probability $\frac{M}{M+N}$. Thus $\frac{m}{m+n}$ appears here in two forms: (1) as an empirical relative frequency and (2) as an *a priori* probability, and $\frac{m'}{m'+n'}$ may be considered as an approximation to $\frac{m}{m+n}$ as well as an approximation to $\frac{M}{M+N}$. Furthermore, unless all the facts are known to us, we shall not be able to differentiate the one case from the other.

May not X itself in our equation $EX' = X$ be a random variable and possess a mathematical expectation EX? Might we not reason as follows in connection with our previously given calculation of the yield of corn in 1930. The method of sampling which we applied gives us a yield A' as the empirical approximation to

[15] Anderson, Oskar, "Korrelationsrechnung," pages 72-80 and 117-122.

the *true* yield *A?* This latter is itself influenced by two groups of factors:

1. Year to year changes such as changes in rural population and in the cultural status of the rural population, changes in the system of crop rotation, increases in acreage, improved technical processes and so forth, resulting in a tendency (I do not want to say "trend") toward a certain steady development of the progressing series of *A's.*

2. There are a number of more or less sudden, irregularly appearing influences, which result in more or less accidental fluctuations. Climatic conditions, shifts in acreage due to the development of changing price relationships and other similar factors fall into this group.

Considered in this light, it would be possible to discern in the magnitude A, for each year, the elements r and n, so that,

$$A = r + n$$

If the law of the distribution of n is such that $En = 0$, then $EA = r$. It is left to the investigator to decide when and in what sense his A' may be taken as an approximation to A and when as an approximation to r. I certainly do not wish to infer that such an extension of the conception of mathematical expectation is always justified. This is a question of economic theory where the statistician is forced to stay in the background.

It is a fact which is more and more being recognized that one cannot reach a knowledge of the "mass phenomena of price formation on the market" by using the methods of the Lausanne School of economics (Lexis).[16] The many millions of equations which are necessary for characterizing every formation of price on the market are absolutely impossible to solve in practice. The American economists (as Henry Moore and his school) admit quite rightly that if the changes of the single variables can no longer be recognized, then the examination of the average relations of the variables is the only way out. In all sciences, the change from the functional to the stochastic mode of thinking is unavoidable if the phenomena to be examined have to be summarized on account of the impossibility of grasping the individual relationships. By determining the correlation coefficient for the average dependency of the single factors, the problem is solved. The characteristic average relationship is substituted for the sys-

[16] See the very convincing paper of Altschul "Die Mathematik in der Wirtschaftsdynamik," cited above, pages 523-538.

tem of simultaneous equations which were supposed to express the existing interrelations. Relationships with different meanings, comprehensible only through the average, or as we now say, stochastic relationships, replaced the functional relationships having a single meaning. Logically, therefore, the calculation of probability had to replace the differential calculation (of the Lausannians).[17] It is a pity that in transferring the correlation method from the field of biometry to the field of economic research, certain facts have been overlooked, as for example, the fact that most economic series are ordered in time. There is much work yet to be done in adapting the correlation method to the problems of economic research, and in devising new methods of approach to such problems. As I tried to show in my book, the many problems can be solved only by introducing the conception of mathematical expectation.

I should like to deal more fully with this subject, but time will not permit. Suffice it to say that my investigations have led to the conclusion that the empirical correlation coefficient of two time series is really a function of three heterogeneous correlation coefficients:

1. The correlation coefficient of the deviations of the individual items of each series from their mathematical expectations.
2. The correlation coefficients of the arithmetic means of these deviations.
3. The correlation coefficients of the mathematical expectations of the individual items themselves.

To my mind this fact explains most of the "nonsense" correlations frequently worked out in economic investigations. Nonsense correlations can easily be avoided through a more careful and clean-cut analysis of the theoretical problems involved prior to making the determination. In every instance the greatest precaution should be taken in applying correlation methods in the field of economic research.

I should like to introduce, briefly, the question of the relation between correlation and causation. This question should be of interest to every student of economic theory. This relationship in Yule's method of deriving the correlation coefficient, which is the method generally presented in most text-books, has been rather neglected. In this connection the derivation of the correlation coefficient which starts from the coefficient of total determination

[17] Altschul, *ibid.*, 528.

seems to me to be more adequately treated.[18] This idea, with the reconstructions necessary for its translation into the language of "stochastic statistics," is as follows:

When studying the causal relation between two observed phenomena, as for example, the relation between air pressure and the temperature at which a certain liquid boils, we may discern in the numerical expressions x and y of our observations, two different elements; (1) those which are really causally connected with each other (which we shall designate as ξ and ψ), and (2) those which may be regarded as "accidental disturbances" or errors of observation (designated as e and ε). We may say, therefore, that:

$$x = \xi + e$$
$$y = \psi + \epsilon$$

We assume that the relation between ξ and ψ cannot be changed and that it is fixed in form. From the standpoint of mathematical analysis, therefore, the relation may be treated as functional and presented in the well known form $\psi = f(\xi)$, or $\xi = f(\psi)$. It is therefore quite reasonable to assume that e is stochastically independent of ξ or ψ and ε of ψ or ξ. We could go even further and postulate the mutual independence of e and ε.

The physicist, who has the experimental method at his command, is usually able to carry out his experiment in such a way that the errors e and ε are very small and consequently do not overshadow the relation between ξ and ψ, and may, therefore, almost be neglected. Consequently he concentrates his attention on the search for the law of the relation between ξ and ψ, that is on the form and constants of the function $\psi = f(\xi)$. The economist, the biologist, and the meterologist are, as we know, not so fortunate as the physicist since on the one hand they are not free to experiment, and on the other they are dealing with those phenomena where the relative values of e and ε are very considerable and furthermore cannot be eliminated, so that outwardly there appears to be little relation between x and y. As an example might be cited the relation between the harvest in northern Bulgaria and wheat exports from Varna Harbor, where, in the first place, the world

[18] Wright, Sewall, "Correlation and Causation," Journal of Agricultural Research, 20, 557-585, 1921. I cite this reference from the paper of Ralph F. Watkins, "The Use of Coefficients of Net Determination in Testing the Economic Validity of Correlation Results," Journal of the American Statistical Association, Vol. XXV, New Series, No. 170, June, 1930, pages 191-197.

price situation appears as a "disturbance." This leads to two problems; (1) not only to find the form of the function $\psi = f(\xi)$ but also (2) to estimate or measure how far this functional (and causal) relation between ξ and ψ is really manifested, and how far it is overshadowed by the disturbances e and ε, or in other words to determine the real degree of association between x and y.

To construct a rational measure of the degree of association between x and y, we shall first consider the most simple case where, with a single pair of observations, the component e is missing altogether, so that $x = \xi$ and $y = \psi + \varepsilon$. It is evident that here the degree of association may very well be expressed by the quotient,

$$\frac{\psi}{y} = \frac{\psi}{\psi+\epsilon} = 1 - \frac{\epsilon}{y}.$$

If, however, the component e is contained in x, a single rational measure of the degree of association is found in the product,

$$\frac{\xi}{x} \times \frac{\psi}{y} = \frac{\xi}{\xi+e} \times \frac{\psi}{\psi+\epsilon} = \left(1 - \frac{e}{x}\right)\left(1 - \frac{\epsilon}{y}\right)$$

It is only necessary to insert the "errors" e and ε into the formula according to their absolute size (always with a plus sign). The maximum value of the product is *1*, and it will be obtained only when e and ε do not exist at all. The minimum value is *0*, which will be obtained only in case the component ξ or ψ or both are missing. Our measure however has a great defect, in that x and y can in no way be obtained from a single pair of observations, and we must therefore use series of such pairs of observations (the more so since a certain number of observations are always necessary in determining the constants of the function $\psi = f(\xi)$.)

Let two series be given, x_1, x_2, x_3, ... x_n and y_1, y_2, y_3, ... y_n, which we assume to be composed as follows:

$$\begin{array}{llllll}
x_1 = \xi_1 + e_1 & & & & & y_1 = \psi_1 + \varepsilon_1 \\
x_2 = \xi_2 + e_2 & & & & & y_2 = \psi_2 + \varepsilon_2 \\
x_3 = \xi_3 + e_3 & & & & & y_3 = \psi_3 + \varepsilon_3 \\
\cdot \quad \cdot \quad \cdot & & & & & \cdot \quad \cdot \quad \cdot \\
\cdot \quad \cdot \quad \cdot & & & & & \cdot \quad \cdot \quad \cdot \\
\cdot \quad \cdot \quad \cdot & & & & & \cdot \quad \cdot \quad \cdot \\
\cdot \quad \cdot \quad \cdot & & & & & \cdot \quad \cdot \quad \cdot \\
x_n = \xi_n + e_n & & & & & y_n = \psi_n + \varepsilon_n
\end{array}$$

Here again, as above, $\psi_i = f(\xi_i)$. The real values of the components ξ, e, ψ, and ε are of course unknown to us, and they cannot be obtained from the above system of equations since the number of the variables is exactly twice as large as the number of equations. The question is, how to measure the degree of association between x and y. In order to proceed we must make several assumptions which, so to speak, take the place of the missing equations. In the first place it is clear that the form and the constants of the function f which connect ξ_1 with ψ_1, ξ_2 with ψ_2, ξ_3 with ψ_3 and so on, must in every case remain the same, since otherwise the first problem, namely that of discovering the causal relation between ξ and ψ could not be solved, and the series x and y would have to be considered as not homogeneous. Furthermore, it is evident that since there is no connection between ξ_i and e_i and between ψ_i and ε_i, the single product $\frac{\xi_i}{x_i} \times \frac{\psi_i}{y_i}$ cannot remain constant for different values of i. In order to arrive at a constant, the individual items ξ_i, x_i, ψ_i, and y_i, must be replaced by such *a priori* characteristics as are common to all members of each series. This again is possible only by introducing additional assumptions. Several such assumptions might be possible. The two which are relatively most important are as follows: (a) It may be assumed that the series of ξ, of e, of ψ, and of ε, are homogeneous; that is to say that the "moments" of the single members of each series (mathematical expectation, standard deviation, and so forth) remain constant. This would represent a simple case, and the one most easily treated mathematically. (b) It may be assumed that the component ξ_i coincides with the mathematical expectation of x_i, and similarly that ψ_i corresponds with the mathematical expectation of y_i. The mathematical expectations of the single items in each series, in this case, need not be constant. It is just here that the further application of the conception of mathematical expectation in economic research offers interesting possibilities.

. The export of corn from Varna Harbor, previously referred to, may be used by way of illustrating the second assumption. There is evidently a close relation between the corn crop of northern

Bulgaria and the quantities of corn exported from Varna Harbor. The exportable surplus A' is derived from the crop statistics by deducting the seed requirements, the amount required for home consumption and the reserves. The figure A' is, of course, subject to a considerable error e, inherent in estimates based upon representative selections. We can, therefore, write $A' = A + e$, where, as shown above, under certain conditions, A may be assumed as being equal to EA'.[19] There is a whole group of factors such as the relative distances of producers from Varna and from other export points in Bulgaria, trade organization, shipping conditions and so forth, which, if acting alone, would cause a certain percentage p of the exportable surplus to be exported *via* Varna. There is a second group of factors such as the world price situation, climatic influences, experience resulting from the liquidation of last year's crop, crop prospects for the coming year, and so forth which appear as disturbances of the percentage p and modify the actual amount exported. By designating the amount actually exported during a given year as B' and the disturbances as ε, we get the following time series:

$$\begin{array}{ll}
A'_1 = EA'_1 + e_1 & \qquad B'_1 = p\,EA'_1 + \varepsilon_1 \\
A'_2 = EA'_2 + e_2 & \qquad B'_2 = p\,EA'_2 + \varepsilon_2 \\
A'_3 = EA'_3 + e_3 & \qquad B'_3 = p\,EA'_3 + \varepsilon_3 \\
\ldots\ldots\ldots\ldots & \qquad \ldots\ldots\ldots\ldots \\
\ldots\ldots\ldots\ldots & \qquad \ldots\ldots\ldots\ldots \\
\ldots\ldots\ldots\ldots & \qquad \ldots\ldots\ldots\ldots \\
\ldots\ldots\ldots\ldots & \qquad \ldots\ldots\ldots\ldots
\end{array}$$

It is also evident that by assuming the series ε to possess a certain law of distribution such as to result in the equation $E\varepsilon = 0$, we would also have:

$$EB'_i = bEA'_i$$

Thus a very interesting case follows from our assumption that a stochastic relationship exists between A' and B', namely, that a functional or causal relationship exists between their mathematical expectations. This can, however, also easily be analyzed on the

[19] Strictly $EA^1 = A + Ee$, but as can be easily shown, Ee will vanish in the calculation of the standard deviation, correlation coefficient, etc., if the deviations are taken from the arithmetic mean.

basis of the procedure worked out in my book with the help of the calculus of mathematical expectations.[20] In order to avoid a misunderstanding, however, I must emphasize that the case (b) represents only one of the different possible hypotheses, the validity of which has to be proved in every individual case. Moreover, I agree with Lorenz that the fertility of the hypothesis ought to be shown by several examples.

We now return to our measure of the degree of association between the two series. It would not be a rational procedure simply to replace the quantities in the product $\frac{\xi_i}{x_i} \times \frac{\psi_i}{y_i}$ by their mathematical expectations, since irrespective of this we would always obtain the maximum value, $+1$, when the mathematical expectations of e or ε equal 0, although this represents the approximately normal case and although each e_i or ε_i, according to their absolute size, considerably exceeds even the values ξ_i and ψ_i. The simplest *a priori* characteristic of our series which takes into consideration also the variability of the series, appears, therefore, to be the standard deviation.[21]

Suppose that σ_ξ, σ_e, σ_ψ, and σ_ε, represent the standard deviations of ξ, e, ψ, and ε respectively. Thus our measure of the degree of association between x and y assumes the following form:

$$H = \frac{\sigma_\xi}{\sigma_x} \times \frac{\sigma_\psi}{\sigma_y}$$

If we write $q_1 = \frac{\sigma_\xi}{\sigma_x}$ and $q_2 = \frac{\sigma_\psi}{\sigma_y}$

then, $H = q_1 \times q_2$

This coefficient, as it relates to the theory of probability, is an *a priori* one, and the question arises as to how to find an empirical

[20] Anderson, Oskar, Korrelationsrechnung, Chap. IV.

[21] The standard deviation translated into the language of mathematical expectations is:

$$\sigma_x^2 = E(x - Ex)^2$$

Within the meaning of the theory of probability σ_x^2 is therefore an *a priori* quantity.

approximation to it. We regard here as the empirical approximation or assumed value, that function of the empirically given series $x_1, x_2, x_3 \ldots\ldots\ldots$ and $y_1, y_2, y_3 \ldots\ldots\ldots$, the mathematical expectation of which (or at least the stochastic asymptote of which) gives the value of q_1q_2. With the aid of the method of mathematical expectations it can easily be proved that such an approximation to H is the product of the two empirical correlation ratios of Karl Pearson:

$$\eta'_{X/Y} \times \eta\ _{/X}$$

Unfortunately these almost always give the marginal value $+1$ for time series, and are therefore of little use to us. There are, however, other methods which we do not have time to discuss.

A very interesting case arises if (on the basis of experience or theoretical considerations) we are justified in assuming, as in our example with the export of corn, that the functional relation between ξ_i and ψ_i is linear, so that $\psi_i = a + b\ \xi_i$, where i may have any value. It is then easily proved that the absolute value of the empirical correlation coefficient r'_{xy} is the assumed value of H, namely,

$$(As_B\ Er^1{}_{xy}) = H = q_1q_2$$

This simple relation is valid, however, only when the functional relation between ξ and ψ is a linear one, and only if all the other assumptions made in our procedure prove to be correct. I believe that this theoretical construction may be of great practical importance in the future insofar as economic prognosis is concerned. Assume, for example, the possibility of proving that the correlation coefficient between the quantity of money in circulation in Bulgaria, and the index of wholesale prices two months later is 0.87. If here, $q_1 = q_2$, it would mean that a linear function of $\sqrt{0.87}$ or 93 per cent exists. If q_1 or q_2 equals *1*, a functional relation of *87* per cent would still exist. And, on the basis of the determination of the amount of money in circulation, one could, with a certain margin of error, which is easily calculated, estimate the wholesale index of prices two months later. The "coefficient of total determination" can also be easily generalized in the case of the so-called multiple correlation.

In conclusion I should like to emphasize that in economic research based on economic statistics, the theory of probability, now neglected in America and in Germany, should be introduced or rather reintroduced. Thereby the limits of probable error could be estimated and many disillusions avoided. This would also lead to more caution being exercised, and to cleaner cut results. Americans are enthusiastic about producing on a large scale, even in statistics. They should recognize the very real danger of producing large errors.