THE INTERPRETATION OF CERTAIN REGRESSION METHODS AND THEIR USE IN BIOLOGICAL AND INDUSTRIAL RESEARCH[1]

By C. Eisenhart

1. **Introduction.** Just as the scientific theorist depends upon the research worker for the facts upon which to build his theory, so does the practical man rely upon empirical relationships to help him estimate (or predict) the value of one quantity from that of another. Sometimes he is interested in assessing the value of some quantity which it is impracticable or impossible to observe directly in a given instance, the estimation being performed with the aid of a previously established relationship between the quantity whose value is sought and another whose value can be determined directly. In other instances he wishes to make use of the relationship existing between two or more quantities to help him adopt a course of action which has a good chance of leading him to a desired result. An example is that of a manufacturer who wishes to exercise control at various stages of a manufacturing process so as to produce a product whose quality lies within a specified range.

In appealing to the interests of the practical man, proponents of statistical methods have often illustrated their writings with beautiful examples of the power of this implement of research, without adequately discussing the abstract ideas that underlie the methods they have promoted—ideas essential to correct statistical thinking. The result has been that to many research workers certain problems with similar objectives appear amenable to identical statistical solution, when in fact intrinsic differences exist which alter considerably the details of their solution.

Such misinformation is particularly prevalent among those whose knowledge of the mathematics of correlation, and of curve fitting, has been derived from the treatment in elementary statistics courses of problems in which no one of the variables stands out from the rest as being *the* dependent variable, with its values determined (not exactly, but within limits) from the values that happen to be assumed by the other variables in the data under investigation. In elementary courses the usual procedure in such cases is to *take* one of the variables as the dependent variable, and then *consider* the others as independent variables. Furthermore, the curve-fitting procedure usually adopted depends on the additional assumption that the values of the independent variables are known exactly (without error)—an assumption often passed by without mention, and one that

[1] Revised from an expository paper presented, under different title, to the American Statistical Association, at Detroit, December 29, 1938, at the invitation of the program committee of the Biometrics Section.

introduces artificiality into the analysis and imposes limitations on the range of applicability of the inferences drawn. This simplification of problems without explicit mention of the fact, fosters misconceptions that are carried over into analyses of data in which the dependent variable is definitely a particular one of the variables and no other—a particularly bad misconception being that the variable whose value is to be estimated automatically assumes the rôle of the independent variable. The calculation and use of dosage-response curves in problems of biological assay constitute an example, and a case which has been correctly solved. The dosage-response curve should be evaluated from a series of observations, with dosage as the *independent* variable, and the curve then used to estimate unknown dosages from observable responses.

It is one object of the present paper to pass in review some of the ideas involved in current curve-fitting practices so that the reader can see for himself why, when one is interested in estimating X from Y, in some instances it is necessary to follow out curve-fitting practices with Y as the dependent variable, and then use the inverse of the relation found. In addition, it is an object of this paper to indicate the types of problem to which this method of inverse regression affords a solution, and to emphasize the confidence interval nature of the estimates it provides. The method will be exemplified by working out in detail a problem arising in the manufacture of cheese, and also a problem concerned with the biological assay of a hormone substance.[2]

2. Mathematical Aspects of the Formulation Of Empirical Relationships. Probably the most obvious way of investigating whether any relationship exists between two variables is that of plotting the observed pairs of values on graph paper. For simplicity we shall confine our attention in this paper to the case of only two variables. While the general trend of the plotted points may suggest the existence of a relationship, the plotted points themselves do not give a definite expression of that relationship, and it is often desirable to have a formula of some sort that expresses it concisely. Furthermore, in all branches of science the data of the research worker are subject to all sorts of fluctuations which tend to make the observational points scatter about a general trend in a band not unlike the Milky Way. Consequently various methods have been developed for inferring from the observations the 'true relation' between the quantities concerned, or, more exactly, a relation which it is hoped will be sufficiently close to the 'true relation' *for the purposes in mind.*

In the development of these methods two rather different viewpoints had to be taken into consideration: first, that of the physical scientist who views the irregular fluctuations as being quite apart from the phenomena under observation and arising solely from inaccuracies of measurement and experimental

[2] Those who are primarily interested in problems of biological assay will find additional material in references [26] to [31]; those whose interests lie in the direction of quality control are referred to W. A. Shewhart [9], and E. S. Pearson [5]. Numbers in [ ] refer to the references at end of the paper.

technique; secondly, that of the biological and social scientists who attribute a large portion of the apparent irregularity of their observations to a real variability which is an essential part of the phenomena studied. That two such divergent viewpoints could be brought together on a common ground is a tribute to the pioneers in mathematical statistics, and the manner in which it has been effected is indicated by the following entry in E. S. Pearson's notebook[3] for 1921–22:

"The purpose of the mathematical theory of statistics is to deal with the relationship between 2 or more variable quantities, without assuming that one is a single-valued mathematical function of the rest. The statistician does not think that a certain x will produce a single-valued y; not a causative relation but a correlation. The relationship between x and y will be somewhere within a zone and we have to work out the probability that the point (x, y) will lie in different parts of that zone. The physicist is limited and shrinks the zone into a line. Our treatment will fit all the vagueness of biology, sociology, etc. A very wide science."

When viewed from this angle, the fundamental problem in the determination of a relationship between two variables, say X and Y, is to determine as accurately as possible from the data in hand the simultaneous probability distribution of the observable quantities, say x and y, considered as random variables. There is, however, a subtle but important distinction between the cases in which the random variability of x and y is due to errors of measurement, etc., and the cases in which this random variability is, as in biological variation, a part of the phenomena under investigation. In the latter we postulate the existence of a probability distribution of the random point (x, y) about some point of location $(\bar{\bar{X}}, \bar{\bar{Y}})$, where the exact meaning of the coördinates $\bar{\bar{X}}$ and $\bar{\bar{Y}}$ depends on the nature of the probability distribution, although they will generally be the coördinates of the mode. In these cases, since (x, y) is subject to biological variation *only*, (x, y) will lie on the line $X' = \text{constant}$ only in cases where $x = X'$. Accordingly, along a line $x = X'$ we shall have the probability distribution of the random point (X', y) about some point of location $(X', \bar{\bar{Y}}_{X'})$. This may not be true when x is also or alone subject to experimental error, for here we postulate a separate probability distribution of (x, y) for each 'true point' (X, Y), and when there are 'errors' in both coördinates (x, y) can lie on the line $x = X'$ when $X \neq X'$. In these cases, the observed distribution of (x, y) for $x = X'$ may result from sampling more than one probability distribution and cannot be interpreted simply. If, however, the X-coördinate is never subject to error, the distribution of (x, y) for $x = X'$ samples the probability distribution of (x, y) for $(X', Y_{X'})$, where $Y_{X'}$ is the true value of Y for $X = X'$. Clearly similar remarks apply in terms of y and Y.

[3] E. S. Pearson, *Biometrika*, vol. XXIX, parts III and IV, (1938) p. 208, writes: "I find on page 1 of my Notes the following statement, which was probably taken down fairly closely from (Karl) Pearson's words: 'The purpose of. . . .' "

Actually at the outset it is not customary to embark on the solution of such a general problem as the determination of the simultaneous probability distribution of x and y. Instead, in the cases where both x and y are subject to 'error', it is customary to assume that the distributions of x about X, and y about Y, are of some particular functional form and then seek to estimate from the data the 'true relation' $\varphi(X, Y) = 0$. Likewise, when x and y are subject only to biological variation, say, it is customary to seek an estimate of the functional relation $\varphi(X, \bar{Y}_X) = 0$, or of the relation $\varphi(\bar{X}_Y, Y) = 0$, where $\bar{Y}_X$ and $\bar{X}_Y$ denote some sort of average (not necessarily the mathematical expectation) of y for a given X and of x for a given Y, respectively, the former being interpreted as being the 'true relation' between X and the average value of y for that X, with a similar interpretation for the latter function. Furthermore, in these cases of mere biological variation it is customary to take $x \equiv X$, $y \equiv Y$, that is, to assume that what we observe are the true values of the quantities, that any errors of measurement are negligible compared to the sampling fluctuations arising from real biological variation.

So far as I know all methods of utilizing observed values of two variables to obtain a relation between the two variables that it is hoped will be sufficiently close to the true relation for the purposes in mind involve the following steps:

(1) To assume that the observational points $(x_1, y_1), (x_2, y_2), \cdots, (x_N, y_N)$ differ from the points $(X_1, Y_1), (X_2, Y_2), \cdots, (X_N, Y_N)$ as the result of observational errors[4] involved in the x, or in the y, or in both coördinates.

(2) To assume, either from the general appearance of the graph of the plotted points or from theoretical considerations, that the relationship between X and Y *is* of the form $\varphi(X, Y; \alpha_0, \alpha_1, \cdots, \alpha_{k-1}) = 0$, where φ is some definite mathematical function involving k, $k \leq N$, constants whose values are unknown. If it is not assumed that φ *is* the true functional relation between X and Y, then it is assumed that the functional relation specified by the φ will be adequate for the purposes in mind.

(3) To choose as an estimate of φ the function $\hat{\varphi} = \varphi(X, Y; a_0, a_1, \cdots, a_{k-1})$ where the a's are those values of the α's that render $\hat{\varphi}$ the function of form φ which is the best fit to the observed points (x_i, y_i), $(i = 1, 2, \cdots, N)$, in some sense of the word "best"; and finally, a step which is too often overlooked.

(4) To carry out some test of goodness of the fit of $\hat{\varphi}$ to the observed points upon the outcome of which rests the decision as to whether a function of the form φ can adequately describe the observed relation between the x's and y's, and, if the decision be affirmative, accepting $\hat{\varphi}$ as an estimate of the true function of form φ.

[4] The word "error" here should be interpreted as "experimental or technical error" from the viewpoint of the physical sciences, (which errors are unbiased in the sense that they average out in the long run), and as "biological variation" from the viewpoint of the biological scientist. In the latter case, if the biological variation is involved in y, and not in x, then $x_i \equiv X_i$ and $Y_i \equiv \bar{Y}_{X_i}$; a similar statement holding if x is in error but not y. In the former case, X and Y are the "true values" of the variables.

In connection with step (4) some results were obtained by W. E. Deming [2] for the case where φ is fit by the method of least squares. He has found that the sum of squared residuals, which is the function to be minimized by the fitting procedure, is fairly sensitive to changes in the functional form of φ, that is, to changes which alter its graph within the range of the observations, but much less sensitive to changes in the values of the parameters involved in a particular functional form. Consequently, by comparing the minimum values of the sums of squared residuals for two different functional forms φ_1 and φ_2 under tentative consideration, it will often be possible to make a good choice between them. On the other hand, it may be possible to alter considerably the values of the parameters in the functional form chosen without appreciably altering the value of the sum of squared residuals. From this it is seen that φ may not be well determined by $\hat{\varphi}$ even though the functional form of φ may be the correct one for the relationship under investigation. For the case where X is exactly known for each observation, with only y subject to error, Deming shows that for the same sum of squared residuals φ is better determined by $\hat{\varphi}$ when there is a long range in X than when there is a short range. In terms of the measure of goodness of fit appropriate to any method of curve fitting these conclusions will probably carry over to that method of curve fitting.

Step (2) also deserves further comment: The function φ may be such that $\hat{\varphi}$ fits the data well within the range of x and y studied, but it must be remembered that an infinite number of other formulae exist which could be adjusted so as to fit the observed points equally well, and some might be found which could be made to fit better. Once a particular functional form for φ has been chosen, if $\hat{\varphi}$ is used to "extrapolate" beyond the range of the observed points, or, if $\hat{\varphi}$ is used as *the* relation between X and Y in any theoretical considerations, it must be remembered that the soundness of any inference that can be made rests to a large extent on the validity of the logic or theoretical considerations that lead to the choice of φ as the expression of the functional relation between the variables, and that the goodness of fit of $\hat{\varphi}$ for one particular batch of data is not a justification of these extensions.

3. Some general remarks on curve fitting practices.

In many cases the assumption is made that a linear relation prevails between X and Y, that is, it is assumed that

$$\alpha_0 + \alpha_1 X + \alpha_2 Y = 0 \tag{1}$$

which may be written in the equivalent forms

$$Y = \alpha + \beta X, \quad \text{where } \alpha = -\alpha_0/\alpha_2\,, \quad \text{and } \beta = -\alpha_1/\alpha_2 \tag{2}$$

$$X = \gamma + \delta Y, \quad \text{where } \gamma = -\alpha_0/\alpha_1\,, \quad \text{and } \delta = -\alpha_2/\alpha_1\,. \tag{3}$$

We are adopting for the moment the viewpoint of the physical scientist, and assuming that (1) represents the true relation between X and Y. We shall return to the case of biological variation later.

A common impression on the part of the research worker, regarding the principles of curve-fitting, seems to be: If one is interested in estimating Y from X, then take $\hat{Y} = a + bX$ as the estimate of (2), and therefore of (1), the a and b being those values which make the line a good fit in terms of the deviations $(y - \hat{Y})$—if one were fitting by the method of least squares one would find the a and b that minimize $\Sigma(y - \hat{Y})^2$, Σ denoting summation over the observed values of y and their corresponding $\hat{Y}$ values; on the other hand, if one is interested in estimating X from Y, then $\hat{X} = b + cY$ is to be fitted, the values of c and d being chosen so as to make $\hat{X}$ a good fit in terms of the deviations $(x - \hat{X})$. *It does not seem to be generally realized that the fitting should be done in terms of the deviations which actually represent "error." Thus when the research worker selects the X values in advance, and holds x to these values without error, and then observes the corresponding y values, the errors are in the y values, so that even if he is interested in using observed values of Y to estimate X, he should nevertheless fit $\hat{Y} = a + bX$ and then use the inverse of this relation to estimate X, i.e. $X = (\hat{Y} - a)/b$, with the best available estimate of Y substituted for $\hat{Y}$.* The situation is quite clear if one approaches the problem from the point of view of fitting the formula to the data with proper attention to which of the variables is in error, as has been recognized for a long time by writers on least squares. If both variables are in error, then this approach also leads to the appropriate solution.[5]

In order to clarify this point it will be helpful to examine the matter a little closer from the viewpoint of the *theory* of least squares.

Let us consider the case where the values of X are selected (or adjusted) by the research worker and then the corresponding values of Y found by observation. So far as the *method* of least squares is concerned in any given instance one *could* minimize $\Sigma(y - \hat{Y})^2$ *and* $\Sigma(x - \hat{X})^2$, thereby obtaining the two lines

$$\hat{Y} = a + bX \tag{4}$$

$$\hat{X} = c + dY, \text{ respectively,} \tag{5}$$

and, unless there existed a perfect correlation between the observed values of X and Y—i.e. unless all of the observed points were exactly collinear, these two fitted lines would differ and yield different estimates of (1). There is nothing in the *method* of least squares to help us choose between these, but from the viewpoint of the *theory* of least squares the correct choice in a given instance is quite clear.[5] The results of the two fitting processes may be given side by side as follows:

[5] See, for example, Deming [1]. Deming pays his respects to a paper by Kummel in *The Analyst* (Des Moines) vol. 6 (1879), pp. 97–105; also to a paper by Uhler, J. Optical Soc. vol. vii (1923), pp. 1043–1066.

$$
(6)\quad \left\{
\begin{array}{ll}
\begin{array}{ll}
\multicolumn{2}{c}{\Sigma(y - \hat{Y})^2 \text{ minimized}} \\[4pt]
\multicolumn{2}{c}{b = \dfrac{\Sigma(x - \bar{x})(y - \bar{y})}{\Sigma(x - \bar{x})^2}} \\[8pt]
\multicolumn{2}{c}{a = \bar{y} - b\bar{x}} \\[8pt]
\textit{Analysis of Variance I} & \textit{d.f.} \\
\text{Total variability of } y\text{'s} & \\
\text{about their mean: } \Sigma(y - \bar{y})^2 & N - 1 \\[4pt]
\text{Reduction effected by (4):} & \\
b\Sigma(x - \bar{x})(y - \bar{y}) & 1 \\[4pt]
\text{Deviations about } \hat{Y}: & \\[4pt]
\Sigma(y - \bar{y})^2 - b\Sigma(x - \bar{x})(y - \bar{y}) & \\
\quad = \Sigma(y - \hat{Y})^2 & N - 2
\end{array}
&
\begin{array}{ll}
\multicolumn{2}{c}{\Sigma(x - \hat{X})^2 \text{ minimised}} \\[4pt]
\multicolumn{2}{c}{d = \dfrac{\Sigma(x - \bar{x})(y - \bar{y})}{\Sigma(y - \bar{y})^2}} \\[8pt]
\multicolumn{2}{c}{c = \bar{x} - d\bar{y}} \\[8pt]
\textit{Analysis of Variance II} & \textit{d.f.} \\
\text{Total variability of } x\text{'s} & \\
\text{about their mean: } \Sigma(x - \bar{x})^2 & N - 1 \\[4pt]
\text{Reduction effected by (5):} & \\
d\Sigma(x - \bar{x})(y - \bar{y}) & 1 \\[4pt]
\text{Deviations about } \hat{X}: & \\[4pt]
\Sigma(x - \bar{x})^2 - d\Sigma(x - \bar{x})\ (y - \bar{y}) & \\
\quad = \Sigma(x - \hat{X})^2 & N - 2
\end{array}
\end{array}
\right.
$$

In all instances Σ denotes summation of the expression following it over all the observed values; $\bar{x} = (1/N)\Sigma x$, the arithmetic mean of the *chosen* values of X; and $\bar{y} = (1/N)\Sigma y$ the arithmetic mean of the *observed* values of Y. The expression in the middle row of each table of the analysis of variance is an immediate consequence of the minimizing process employed; the last row is obtained by subtraction.

Let us now interpret these analysis of variance tables. *On the left,* $\Sigma(y - \bar{y})^2$ gives a measure of the observed variability of the y values, a portion of this variability being due, we suppose, to the dependence of Y on X. The second row of table I gives the portion (the maximum portion on the basis of the observations) of the observed variability of the y's that can be attributed to the dependence of Y on X, and the last row indicates the magnitude of the rest, that is, the magnitude of the portion of $\Sigma(y - \bar{y})^2$ that must be attributed to "error" (and, this portion has been minimized by the fitting process). *In short, remembering that we are dealing with the case in which the values of X are chosen by the research worker and only the values of Y are subject to error, the relation between X and Y being as in* (1) *or its equivalent form* (2), *we see that the analysis of variance table on the left separates $\Sigma(y - \bar{y})^2$ into portions whose meanings are clear.* In particular, since unrelated variables can exhibit relationship in finite samples, the test of whether β is really different from zero resolves itself into examining whether the variance ratio

$$
\left(\frac{b\Sigma(x - \bar{x})(y - \bar{y})}{1}\right) \bigg/ \left(\frac{\Sigma(y - \bar{y})^2 - b\Sigma(x - \bar{x})(y - \bar{y})}{N - 2}\right)
$$

is of a magnitude that *may* be taken to indicate $\beta \neq 0$ in the sense that the risk of falsely rejecting the hypothesis that $\beta = 0$ by so doing is of an acceptable smallness.

The analysis of variance table *on the right,* on the other hand, can be misleading if it is interpreted hastily. In the first place $\Sigma(x - \bar{x})^2$ represents the variability in the chosen values of X which resulted from the way in which the research worker selected (or adjusted) them, and it is to be noted that the corresponding values observed for Y have in no way entered into their determination. Consequently the *apparent* dependence of the x on the y, measured by d, or more effectively by the second row of table II, is a spurious dependence, and the last row of this table cannot be interpreted as being a measure of the "error" in the x values, in the sense of being that portion of the variability of the x values which cannot be accounted for by the variability of the y values. *Briefly stated, when the values of x have been selected by the research worker and the corresponding y values observed, the line obtained by minimizing $\Sigma(x - \hat{Y})^2$ is meaningless, and* (4) *is accordingly the only correct estimate of the postulated linear relationship between X and Y, wherefore, if it is desired to reason from Y to X this must be done by means of $X = (\hat{Y} - a)/b$, namely* (4) *solved for X.*

In the preceding paragraphs we have discussed the case where one of the variables is subject to random variation, and the other takes only those values selected (or, to which it is adjusted) by the research worker. Without loss of generality we took Y to be the former variable, and X the latter. Actually we have discussed only the case in which (1), or one of its forms, (2) or (3), is assumed to express the 'true relation' between X and Y. That is, we have been discussing the case where y varies about Y as a result of experimental 'error,' and we have not treated the case where y is subject to biological variation.

If X takes only those values selected by the research worker, and y is subject to biological variation but is known without observational error, so that $y = Y$, (1) no longer applies for the reasons given in section 2, but it must be replaced by

$$(7) \qquad \alpha_0 + \alpha_1 X + \alpha_2 \bar{Y}_X = 0$$

where $\bar{Y}_X$ is the 'average' value (but not necessarily the arithmetic mean or mathematical expectation) of Y for the value of X denoted by the subscript. Clearly (7) may also be written in a form corresponding to (2),

$$(8) \qquad \bar{Y}_X = \alpha + \beta X, \quad \text{with} \quad \alpha = -\alpha_0/\alpha_2 \quad \text{and} \quad \beta = -\alpha_1/\alpha_2$$

or in a form corresponding to (3),

$$(9) \qquad X = (\bar{Y}_X - \alpha)/\beta = -\alpha/\beta + (1/\beta)\bar{Y}_X .$$

With this latter form we may contrast

$$(10) \qquad \bar{X}_Y = \gamma + \delta Y$$

a relation expressing "the true average value of X for a given Y" as a linear function of Y. Equation (10) is of interest, as well as (8), when X is free to vary in samples according to the biological variation associated with it, but when the distribution of values of X is dictated by the wishes of the research worker, as in the case under discussion, it can be demonstrated that (10) is of no value for purposes of inference.

The method adopted for estimating (7), or one of its alternative forms, will depend on what "average" $\bar{Y}_X$ is taken to be. If, as is usually the case, $\bar{Y}_X$ denotes the true arithmetic mean (or mathematical expectation) of Y for a given value of X, then (4) fitted by the method of least squares as above affords an unbiased estimate of (8). Or, if $\bar{Y}_X$ were taken to be the true median of Y for a given X, then in general one would fit (4) by minimizing $\Sigma \mid y - \hat{Y} \mid$, the summation being taken over the observed y values. As in the discussion of the case involving experimental error, to estimate X from Y one would estimate (9) with (4) solved for X, and in a particular instance replace $\hat{Y}$ by the best available estimate of $\bar{Y}_X$ from the data in hand. This brings out the strong similarity between statistical procedures appropriate when the variables are subject to experimental error and when on the other hand they are subject to biological variation but can be accurately observed.

A great injustice would be done to many previous writers by failure to mention at this point that the ideas and the conclusions reached in the preceding paragraphs have been appreciated for a long time by some of the writers who have developed the theory and applications of curve fitting. At most, the preceding paragraphs are but an emphatic way of presenting what these experts would regard as obvious.

4. Effect of Limiting the Range of Either Variable in the Sampling Process.

In the preceding section we have discussed the situation in which one of the variables does not vary at random, but assumes only those values selected by the research workers. We have seen that in such cases this variable must be taken as the independent variable in applying any curve-fitting procedures. The same conclusion applies when both of the variables are subject to biological variation but the sampling process limits the observed range of one of the variables—only the results obtained by using the restricted variable as independent variable can be expected to give an unbiased description of the underlying relationship in the population sampled. If X is the variable for which the range of observable values is constricted by the sampling process, this means that the relation (8), for the population sampled, can be estimated from the data; but relation (10) for the population is unattainable.

To illustrate this point it will be sufficient for our purposes to consider Figure 1 which has been constructed from some artificial data which are especially suited to this purpose. We shall suppose that Y is the dependent variable and X the independent variable, and that the complete array of points shown arose from a sampling process in which neither X nor Y was restricted. It will be noticed that the observational points lie in a band sloping upward to the right and that as x increases by one unit the distribution of the corresponding y's moves up by one-half a unit. We may consider the points of the entire band shown as portraying the relationship between X and Y in the large, that is, when a point (x, y) is selected at random without restrictions on either X or Y. The slanting line labelled (I) indicates the "average" relationship prevailing between Y and

X, that is, for a given value of X the arithmetic mean of the corresponding observed values of Y is given by the point on this line with abscissa X.

Let us now consider the situation in which the points have been selected with restriction on X. As the results of such a procedure of selection let us take only those points between the two vertical lines drawn just to the right of $X = 3$ and just to the left of $X = 7$. It will be seen that this does not upset the average y for a given value of x within the prescribed limits, i.e. $\bar{Y}_X$ is unaltered for $3 < X < 7$. *In other words, the introduction of a restr'ction with regard to X,* the independent variable, *has not spoiled the inferences with regard to Y, when Y is considered as the dependent variable*—that is, when we are arguing from X to Y.

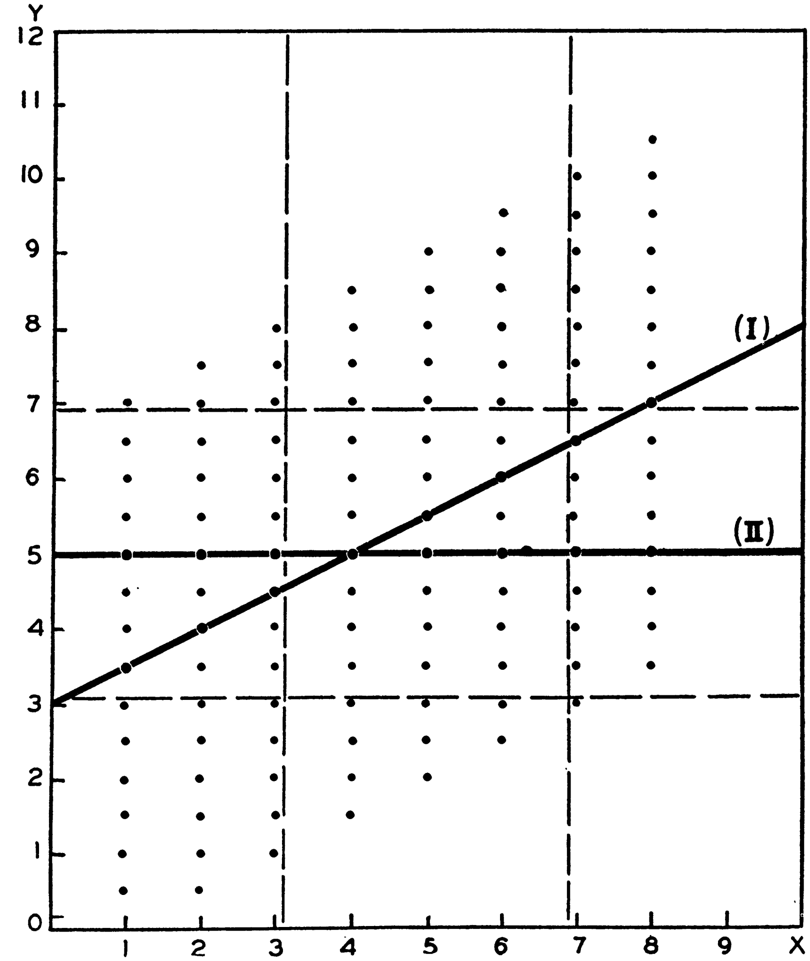

FIG. 1

Consider now the effect of restricting the observed y in a sampling process and then trying to infer about $\bar{Y}_X$ in the population at large from given values of X. In Figure 1 this corresponds to considering, say, only those points that lie between the horizontal lines just above $Y = 3$ and just below $Y = 7$. It is seen immediately that in this case, i.e., between the *horizontal* lines, for every value of X the average of the observed Y values is $Y = 5$, and consequently the relation of Y to X is portrayed by the line numbered (II). *It is seen that in this case the "apparent" relation is not the correct one. Accordingly, we conclude that the restriction of the dependent variable is liable to seriously distort the relationship, so that what is observed is not representative of the true underlying situation.*

The demonstration that we have chosen is simple and artificial but the conclusions drawn apply in general, namely, the restriction of X does not alter the regression of Y on X, but the restriction of Y does. For further illustrations and a very readable discussion see Chapter 19 of *Methods of Correlation Analysis* by Mordecai Ezekiel.

As a special case of a situation in which the "observed" y's are restricted in some way or other we may turn the problem around and note the limiting case where Y is not a random variable at all but is given certain assigned values by the research worker and the corresponding values of X are ascertained by observation. It is evident from what has gone before that in such a case any formula that expresses the average value of y for a given value of x for the data thus collected is useless for inferring anything about the average value of Y for a given value of X in the "population" at large.

5. **Variables Subject to Biological Variation and also to Errors of Observation.** In the preceding paragraphs we have been supposing that the variables were subject either to errors of measurement, or to biological variation, but we excluded the case in which both types of variation were in operation simultaneously. It is reasonable to suppose that errors of measurement are present in biological work just as they are in the physical sciences, though it will usually be found that the variability between biological specimens is far greater than the maximum variability that could be attributed to errors of measurement. Accordingly, in most biological work true biological variations force errors of measurement into the background. It is usually possible to check up on this by making two or more determinations for each specimen and then comparing the variation between determinations with the variation between specimens by means of the analysis of variance technique developed by R. A. Fisher [3]. When only one determination is made per specimen the two variations cannot be distinguished.

Even if observational errors are in the background, it is of importance to know the consequences to be expected when they are superimposed on biological variation. Ezekiel discusses this phase of the subject in detail in chapter 19 of his book mentioned earlier, and a survey of his conclusions in terms of what we have discussed above will be sufficient for our purposes: (*a*) If $\bar{Y}_X$ denotes the average value of Y corresponding to the X denoted by the subscript (in a certain sense of the word "average") and is a linear function (8) of X, then if the X values are free from errors of measurement but the y values are subject to random errors, uncorrelated with the true Y values, and which average out in the long run (in the same sense of "average" as above), then (4) fitted by the method consistent with the meaning of "average" provides an unbiased estimate of (8), in the sense that its "average" value in repeated sampling will be (8), and the effect of the errors of measurement is merely to decrease the precision with which (8) can be estimated from the given set of X values; (*b*) if the situation is as in (*a*) with the exception that the errors are correlated with the true Y values, then not only will their presence affect the precision of (4) as an estimate of (8), but it will render (4) a biased estimate of (8), the tendency being an underestimation of the existing correlation; (*c*) if random errors affect the independent variable correlated or uncorrelated with its true values, then (4) will be an unreliable estimate of (8), and may be markedly biased whether or not the errors of measurement affect the dependent variable; and, if non-random errors of measurement are present they tend to render (4) a more or less unreliable estimate of (8), quite regardless of the variables to which they apply.

The practical significance of these principles in regard to variables subject to biological variations is that if large errors of measurement enter into the determination of some variable, provided these errors are *random* that variable may still be used as the dependent variable without introducing appreciable bias in the estimation equation if enough observations are available to approximately balance out the errors; but any use of that variable as the independent variable will almost surely yield results that understate the actual relationship, and if the errors are not random, they will tend to bias the results quite regardless of the variables affected by them.

6. **An Industrial Problem.** With the preceding discussion in mind let us now direct our attention to a problem that arises in connection with the manufacture of cheese. One of the measures of the quality of a cheese is the percent of fat it contains. In the cheesemaker's notation this is given by the fat-drymatter ratio, F/DM, which is usually written as percent since the fat is contained in the total dry matter. Experience in cheese making has shown that the casein-fat ratio, C/F, of the milk out of which the cheese is made influences the F/DM of the finished cheese, and that the relationship is approximately linear, with a negative slope, for the range of values of these variables usually studied.

Since 45% is the lower limit of F/DM for an acceptable cheese as specified by law, cheese manufacturers are interested in standardizing the C/F ratio of the milk they use, which they can do by separating the milk and cream from individual sources and then putting them together again in proper proportions so that the resulting cheese will have a good chance of meeting the legal requirement *at least.* Figure 2 portrays some results obtained by standardizing the C/F ratio at different values, the individual points representing 149 different batches of cheese manufactured in October, 1936 at a particular factory.[6] It is seen that the relationship prevailing between C/F and F/DM in these data takes the form of a rather wide sloping band and not as a close clustering of points about a well-defined trend.

If a cheese manufacturer is able to infer from data of this sort a reliable answer to a question like the following, he will be able to improve the economic efficiency of his plant: "To what value should C/F be standardized in order that we may expect F/DM to *exceed* 45 in, say, 95% of our future experience?" Unfortunately this type of question, very easy to phrase, is usually exceedingly difficult to answer, and, indeed, the very existence of an answer depends on an assumption of some sort of stability in the manufacturing process, and in the materials used, which enables a future observation to be estimated at least within limits

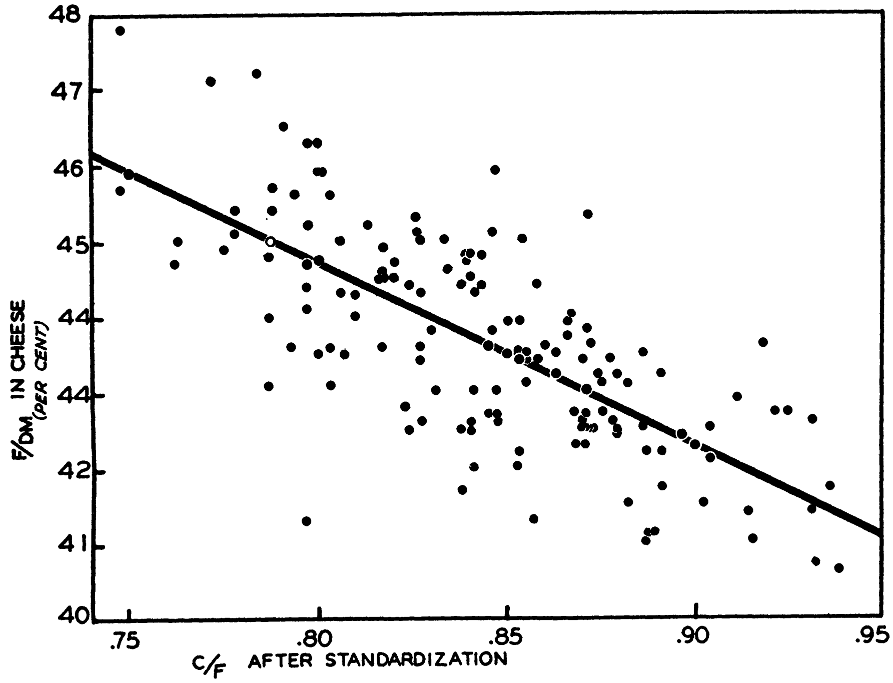

FIG. 2

[6] These data were brought to me by Professor Walter V. Price, of the Department of Dairy Industry of the University of Wisconsin, in connection with a different but related problem, and I wish to acknowledge my gratitude to him for permission to use them in the present discussion. It will be noted that F/DM is given as a per cent, whereas C/F is given as a decimal fraction. This is the customary procedure with dairymen, and arises from the fact that C/F is merely an index involving two different quantities distinguishable in the milk, and cannot be interpreted as a per cent in the same way as the F/DM ratio.

from available experience. In the succeeding paragraphs we shall present a solution that will depend for its applicability upon the following assumptions:

Let Y denote the true F/DM ratio of a finished cheese, X the true C/F ratio in the milk from which it was made, and let $\bar{Y}_X$ denote the true arithmetic mean of Y associated with the value of X indicated by the subscript.

Assumption I: We shall assume that the dependence of $\bar{Y}_X$ on X is linear and given by

$$(8') \qquad \bar{Y}_X = \alpha + \beta X = \alpha' + \beta(X - \bar{x}), \quad \text{with} \quad \alpha' = \alpha + \beta\bar{x}$$

where $\bar{x}$ denotes the arithmetic mean of the true C/F values corresponding to the points shown in Figure 2.

It should be noted that $\bar{x}$ and its value do not enter into the specification of the linear relationship but only into the alternative expression of it.

Assumption II: We shall assume that X is determined without error in a given instance, and the differences $(y_X - \bar{Y}_X)$ between the observed values of F/DM, say y_X, and their corresponding mean values, $\bar{Y}_X$, may be regarded as drawn independently at random from a population in which $(y_X - \bar{Y}_X)$ are normally distributed about zero with a variance, $\sigma^2_{Y \cdot X}$, which is the same for all values of X.

Since these assumptions are restrictive it is necessary in connection with a given practical problem to ascertain whether they are acceptable on the available evidence before proceeding to the application to the problem in hand of methods depending on them for validity. Before applying to a problem of his own any of the methods presented in the following paragraphs, the reader should investigate the tenability of these assumptions with regard to his type of data. Methods for examining whether data of a given type exhibit "statistical control" are available in the literature and the reader is referred especially to the writings of W. A. Shewhart [9, 10]. To date experience has shown that it is very difficult to attain and maintain statistical stability in connection with industrial processes. On the other hand, it is uselsss to try to answer questions of inference such as the above until a fair degree of statistical stability is attained, whether statistical processes are employed or not. The success along these lines that has been attained in industry is a great tribute to Shewhart and his insistence on attention to this phase of the application of statistical methods to practical problems. The sooner workers in other fields turn their attention to questions of statistical control, the sooner mathematical statistics will be of some value to them.

From an examination of C/F and F/DM values from the same factory over a period of months it appears that although a relation of the type (8′) above seems to exist in most instances, it is not stable with regard to the values of α and β. Consequently, unless the source of this instability can be discovered and either removed, or allowed for, the answer to the above question is more or less unattainable. In order to exemplify the method, however, we shall proceed as if statistical control were a fact and assumptions I and II tenable.

It is clear, I think, from comments in the early part of this paper that if we let $Y = F/DM$ and $X = C/F$, since the C/F values have been chosen by the cheese makers, we shall have to infer about X from the relation of Y to X, the latter being considered as the *independent variable.* Furthermore, it is a consequence of assumption II that fitting

$$(4') \qquad \hat{Y} = a + bX = a' + b(X - \bar{x})$$

by least squares will provide the most accurate estimates of α and β in (8'). That $a' = \bar{y}$, the arithmetic mean of the observed y values is evident when (4') is compared with (4) and (6). Performing the calculations it was found that

$$(11) \qquad \hat{Y} = 64.38 - 24.58X = 43.63 - 24.58\ (X - .8439),$$

for the data shown in figure 2.

If now we ask "What value of C/F will to the best of our knowledge result in $F/DM = 45$ *on the average* in the future?", the answer is obtained by setting $\hat{Y} = 45$ in (11), solving for X, from which it is found that C/F $(= X)$ should be taken equal to $(64.38 - 45.00)/24.58 = .7884$, and this point is indicated by the black dot with white center on the line in Figure 2. We must remember, however, that (11) is merely *an estimate* of (8'), and that the value of $\hat{Y}$, namely 45, obtained by inserting $X = .7884$ in (11), is merely an estimate of the true $\bar{Y}_{.7884}$, which may not be 45 at all. Indeed the use of $\hat{Y}$ for a particular value of X to estimate the true $\bar{Y}_X$ for that X is mathematically equivalent to the customary procedure of using $\bar{y}$, the mean of all of the observed y to estimate $\bar{Y}$, the true mean of the Y population.

In recent years it has become customary to perform such estimations, not by single value, but by means of confidence intervals, a confidence interval for $\bar{Y}$ being of the form

$$\bar{Y}_1 \leq \bar{Y} \leq \bar{Y}_2$$

where $\bar{Y}_1$ and $\bar{Y}_2$ are functions of the observed values of Y, i.e. of $y_1, y_2, \cdots y_N$, and of the confidence coefficient chosen. If a confidence coefficient, of $1 - \epsilon$ is adopted $(\epsilon > 0)$, then the interpretation of such an inequality is as follows: If inequalities of this form are used whenever it is desired to estimate $\bar{Y}$ from the observed y's, then in the long run we may expect $100\cdot(1 - \epsilon)\%$ of such estimations to be correct, that is, in $100\cdot(1 - \epsilon)\%$ of the cases in which we apply intervals of form (6) they will include $\bar{Y}$ within their limits. Such limits are sometimes referred to as *fiducial limits* and the associated degree of confidence termed the *fiducial probability* of the estimation being correct.[7]

[7] There is an ever-growing literature on this mode of estimation, and a list of references to expository treatments of the subject will be found at the end of the paper together with a few other pertinent references.

From Fisher's 1935 paper it appears that he wishes to restrict the use of the words *fiducial probability, fiducial limits*, etc. to the cases in which a sufficient statistic exists for the parameter tc be estimated. Since he introduced the use of these words in this con-

We shall now show how to set up confidence intervals for $\bar{Y}_X$ in terms of $\hat{Y}$ for that X, and by an extension of the argument, we shall show how to make a probability statement about the difference $(y' - \hat{Y})$ in repeated sampling, where y' is an observation not involved in the evaluation of $\hat{Y}$. The connection of this type of probability statement to the question asked above will be pointed out and its relation to the ideal answer to that question discussed.

In the succeeding paragraphs we shall make use of the following mathematical results:

(A) Assumptions I and II imply that in repeated samples *involving the same values of* X the fitted line $\hat{Y}$ of (4′) will be normally distributed about the true line $\hat{Y}_X$ of (8′) with a variance

$$\sigma_{\hat{Y}}^2 = \sigma_a^2 + (X - \bar{x})^2\sigma_b^2 \tag{12}$$

in which

$$\begin{aligned} \sigma_a^2 &= \sigma_{Y\cdot X}^2/N \\ \sigma_b^2 &= \sigma_{Y\cdot X}^2/\Sigma(X - \bar{x})^2 \end{aligned} \tag{13}$$

where Σ denotes summation over the N actual values of X involved, $\bar{x}$ is the arithmetic mean of these values of X, and $\sigma_{Y\cdot X}^2$ is the true variance of Y for a fixed value of X (and assumed independent of X). The condition that the sampling be confined to the same values of X is an essential part of the statement as can be seen from the original argument by Working and Hotelling [12] which is outlined by Rider [6]. The result is given by Fisher [3] sec. 26.

(B) When $\sigma_{Y\cdot X}^2$ is unknown, a convenient estimate from the sample is

$$s_{y\cdot x}^2 = \Sigma(y - \hat{Y})^2/(N - 2), \tag{14}$$

the distribution of $(N - 2)\, s_{y\cdot x}^2/\sigma_{Y\cdot X}^2$ being as χ^2 with $N - 2$ degrees of freedom and independent of the distribution of $(\hat{Y} - \bar{Y}_X)$.[13]

(C) *Student-Fisher theorem*: The ratio of any quantity d normally distributed about zero with standard deviation σ, to an estimate s having the property that ns^2/σ^2 is distributed *independently* of d as χ^2 with n degrees of freedom, is itself distributed as Student's t for n degrees of freedom.[8]

Letting $S_{\hat{Y}}^2$ denote the estimate of $\sigma_{\hat{Y}}^2$ obtained by substituting $s_{y\cdot x}^2$ for $\sigma_{Y\cdot X}^2$ in the quantities (13), it follows from (A)–(C) that

$$t = \frac{\hat{Y} - \bar{Y}_X}{S_{\hat{Y}}} \tag{15}$$

nection, he has some sort of right to specify their usage. Accordingly Neyman's confidence intervals are of more general availability, and when a sufficient statistic does exist both the fiducial limits and the limits of Neyman's shortest confidence limits (or of his short unbiased confidence intervals) will be found to depend on this sufficient statistic, although the interval between the limits may not be the same in the two cases, Neyman bringing an additional principle into play to assist in the location of his intervals.

[8] Fisher [4]; "Student" [11].

is distributed as Student's t for $N - 2$ degrees of freedom. Consequently if $t_{.05}$ denotes the number for which $P\{|t| > t_{.05}\} = .05$ where t is as in (15), and $|t|$ denotes the numerical value of t, it follows that the probability is .95 that random variations in the y's for the values of X chosen will yield a value of $\hat{Y}$ for which

$$-t_{.05}S_{\hat{Y}} \leq \hat{Y} - \bar{Y}_X \leq +t_{.05}S_{\hat{Y}} \tag{16}$$

is true, that is, a value of $\hat{Y}$ for which

$$\hat{Y} - t_{.05}S_{\hat{Y}} \leq \bar{Y}_X \leq \hat{Y} + t_{.05}S_{\hat{Y}} \tag{17}$$

is true. Accordingly, if we assert *in a given instance* that (17) is true, *there is no way of telling whether our assertion is correct,* but in the long run the $\hat{Y}$'s we calculate from the data we observe may be expected to differ from their $\bar{Y}_X$ values in such manner that (16) will be correct in 95% of our experience, so that we may expect to be correct in 95% of the assertions we make about $\bar{Y}_X$ using (17).

For the data of figure 2 the quantities needed in addition to (11) are

$$\frac{1}{N} = \frac{1}{149} = .00671141 \qquad \Sigma(X - \bar{x})^2 = .274796$$

$$s^2_{y \cdot x} = .9448 \qquad t_{.05} = 1.979, \text{ for 147 degrees of freedom.}$$

For $X = .7884$ it is easy to verify that $(X - .8439)^2 = .0030$, and substituting in (12) with $\sigma^2_{Y \cdot X}$ replaced by $s^2_{y \cdot x}$ gives $S_{\hat{Y}} = .1290$ *for* $X = .7884$, and, since $\hat{Y}$ equals 45 for this value of X, we may assert

$$44.744 \leq \bar{Y}_{.7884} \leq 45.256, \tag{18}$$

and we are correct in this assertion unless a 1 in 20 chance event has occurred. Since these limits do not differ widely from 45, we see that we *may hazard* the prediction that, if $X = C/F$ is standardized to .7884, then the values of $Y = F/DM$ in our future experience will be distributed about a mean fairly close to 45. This prediction is based not only on the assumption that we are sampling a stable statistical population, but also on the presumption that (18) *is* true. $\bar{Y}_{.7884}$ may really lie outside and at quite a distance from this interval. The results of a sampling experiment which illustrate this point in connection with confidence limits for a sample mean will be found in Shewhart [10].

Let us now see how the preceding type of argument may be extended to take into consideration a single additional y ($= F/DM$) value. Let y' denote an additional value of Y not included among those used to construct the regression $\hat{Y}$, and let X' be the value of X to which y' corresponds. If y' be an *independent* observation, then

$$(y' - \bar{Y}_{X'}) \quad \text{and} \quad (\hat{Y}' - \bar{Y}_{X'}),$$

where $\hat{Y}'$ denotes the value of $\hat{Y}$ corresponding to $X = X'$, are normally and independently distributed about zero with variances $\sigma^2_{Y \cdot X}$ and $\sigma^2_{\hat{Y}'}$ respectively.

Since the difference of two quantities normally and independently distributed about zero is also distributed normally about zero with variance equal to the sum of the respective variances, it follows that $(y' - \bar{Y}_{X'}) - (\hat{Y}' - \bar{Y}_{X'}) = (y' - \hat{Y}')$ is normally distributed about zero with the variance $\sigma^2_{y \cdot x} + \sigma^2_{\hat{Y}'}$. Using $s^2_{y \cdot x}$ to estimate $\sigma^2_{y \cdot x}$, which is involved in both of these terms, it follows from (C) that

$$t = \frac{y' - \hat{Y}'}{\sqrt{S^2_{\hat{Y}'} + s^2_{y \cdot x}}}, \tag{19}$$

where $\hat{Y}'$ is the value of (4′) for $X = X'$ and y' is an additional value of Y for $X = X'$ and $S_{\hat{Y}'}$ the value of $S_{\hat{Y}}$ for $X = X'$, is distributed as Student's t for $N - 2$ degrees of freedom. It should be noticed that here the estimate $s^2_{y \cdot x}$ obtained in connection with $\hat{Y}$ carries all of the burden of estimating $\sigma^2_{Y \cdot X}$. Accordingly, unless our *combined experience with regard to y' and $\hat{Y}'$* is such as would occur 1 time in 20, i.e. unless t of (19) numerically exceeds $t_{.05}$ for $N - 2$ degrees of freedom, it follows that

$$-t_{.05}\sqrt{S^2_{\hat{Y}'} + s^2_{y \cdot x}} \leq y' - \hat{Y}' \leq t_{.05}\sqrt{S^2_{\hat{Y}'} + s^2_{y \cdot x}} \tag{20}$$

which may also be written as

$$\hat{Y}' - t_{.05}\sqrt{S^2_{\hat{Y}'} + s^2_{y \cdot x}} \leq y' \leq \hat{Y}' + t_{.05}\sqrt{S^2_{\hat{Y}'} + s^2_{y \cdot x}}. \tag{21}$$

If, therefore, y' denotes a *future* observation, unless our experience to date (contained in $\hat{Y}$ and $S_{\hat{Y}}$) *and* our future experience with regard to y' are such as to make t of (19) exceed $t_{.05}$ numerically—it being supposed we are sampling a statistically stable universe—then if we predict limits for y' by means of (21) we can associate a confidence of .95 with this *combined* procedure—that is, if we make a habit of evaluating regression lines $\hat{Y}$ and of predicting new observations with their aid by means of (21), then in 95% of the cases in which we take *independent paired steps* of this sort we may expect to be correct with regard to our prediction of y'. It should be noted that if $\hat{Y}$ is "away out" in the first place, which may occur by chance, the combined experience of y' and $\hat{Y}'$ will probably be "away out" too, although y' may be near $\bar{Y}_{X'}$ where it belongs. The 95% wager applies to the combined steps of getting $\hat{Y}$ and y' and *not* to the single step laying off an interval about $\hat{Y}$ in hopes of "catching" y'. In consequence one should not keep on using one regression $\hat{Y}$ over and over again, but should be continually amending "experience to date" as data accumulate.[9]

It should be noted that the above procedure does not yield us an interval which may be expected to include 95% of the future values of y. Such a range

[9] H. Working and H. Hotelling discussed this use of regression to forecast future values, but did not, as far as I can see, emphasize the confidence interval nature of the argument, nor the fact that the probability concerned refers to the two steps involved, and not merely to the latter. The same may be said with regard to Schultz's paper [8].

would be an estimate of the range within which 95% of the population values lie. The difficulties attending the estimation of this type of range are discussed by Shewhart [10], and it appears from his work that in the present state of our knowledge very large samples are required for this purpose. In addition, by a beautiful example, Shewhart shows how a failure to distinguish between confidence intervals associated with a given confidence coefficient, say .95, and intervals containing 95% of the population values, can lead to statements which are quite false.

Recalling to mind that we have been going through all of this reasoning with the aim of finding a way of deciding to what value of C/F ($= X$) we should tell the dairyman to standardize his milk if he wishes to produce cheese for which F/DM ($= Y$) is 45 *at least*, we see that our problem consists in getting a lower limit to y' where X' is the value at which we shall advise him to standardize. If, therefore, we leave the right side of the inequality (21) open so that we have

$$\hat{Y}' - t'_{.05}\sqrt{S^2_{\hat{Y}'} + s^2_{y\cdot x}} \leq y', \tag{22}$$

where $t'_{.05}$ is the value of t for which $P\{t < -t'_{.05}\} = .05$, the *sign* of the t value in (19) being considered now, then we seek that value of X, which makes the left side of this equal to 45. For, if y' correspond to this value of X, call it X', then *unless our experience to date plus our future experience* with y' is such as we may expect to occur 1 time in 20 in the long run, y' will be greater than 45, as desired. In other words, we want to solve

$$a + b(X' - \bar{x}) - t'_{.05}\sqrt{s^2_{y\cdot x}\left\{1 + \frac{1}{N} + \frac{(X' - \bar{x})^2}{\Sigma(X - \bar{x})^2}\right\}} = Q \tag{23}$$

for X', where $Q = 45$ in this case. By straightforward algebra the general solution is found to be

$$X' = \bar{x} + \frac{b(Q - a)}{C} \pm \frac{(t'_{.05})s_{y\cdot x}}{C}\sqrt{B(Q - a)^2 + \left(\frac{N + 1}{N}\right)C} \tag{24}$$

in which $a = \bar{y}$, $B = 1/\Sigma(x - \bar{x})^2$, and $C = b^2 - (t'^2_{.05})(s^2_{y\cdot x})(B)$, and the sign before the last term is + if b is positive and − if b is negative.

From the data involved in the present problem $N = 149$, $\bar{x} = .8439$
$a = 43.63$, $b = -24.58$, $B = 3.6391$, $s^2_{y\cdot x} = .9448$, $s_{y\cdot x} = .9720$
and for $t'_{.05} = 1.656$, the one-sided 5% value for 147 degrees of freedom, $C = 594.7479$.

Substituting these values in (24) *we find* $X' = .7207$, *and this is the value to which the dairyman should standardize his* C/F *ratio.* If he does, then *unless* the experience to date, leading to $\hat{Y}$ of (11), and the future experience with regard to any new y ($= F/DM$) value—unless these *combined experiences* are such as to shove the t of (19) beyond the *one-sided* 5% value of t for 147 degrees of freedom *and* in the negative direction, the predicted value of y ($= F/DM$) will be 45 *at least.* In this sense we may have 95% confidence that our prediction will be correct.

It is clear that the preceding solution can be set up for any desired degree of confidence, say $1 - \epsilon$, by choosing t'_ϵ which is the value of t for which $P\{t < -t'_\epsilon\} = \epsilon$ for the degrees of freedom involved. Furthermore, if an upper limit, instead of a lower limit, were desired, the solution would be the same except for an interchanged usage of the + and − signs before the last term of (24)—for an upper limit one would take a − if b were positive and a + if b were negative. For values of Q not too different from $\bar{y}$ it will usually be possible to find the solution corresponding to the level of confidence desired. However, it is quite possible that a solution may not exist for the value of Q desired, if this be too distant from $\bar{y}$. This difficulty will arise whenever $[(N + 1)/N](t'_\epsilon)^2 s^2_{y \cdot x} B$ is larger than $B(Q - \bar{y})^2 + [(N + 1)/N]b^2$, in which case the radical is imaginary, and no real solution of (24) exists. By graphing the left side of (22) for several values of X' the reason why such cases occur can be readily appreciated.

Since the confidence coefficient in reality relates to the difference $(y' - \hat{Y})$ in which both y' and $\hat{Y}$ are random variables, when applying this method to a particular industrial (or other) problem, one should make repeated $\hat{Y}$ estimates of $\bar{Y}_X$ from time to time in order to insure that the $\hat{Y}$ used is not away off from $\bar{Y}_X$. As mentioned earlier $\hat{Y}$ will assess $\bar{Y}_X$ more accurately if the X values used are spread over a rather wide range—this follows from the nature of (12). By frequent determinations of $\hat{Y}$ even better estimates of $\bar{Y}_X$ can be obtained by pooling the data to date, provided no departures from statistical stability are detected. In this way an increasingly reliable estimate of X' can be determined. By standardizing with $X = X'$ and keeping an eye on the resulting y values, one will be able to see whether this choice of X' is operating satisfactorily. Also, and more important probably, by standardizing $X = X'$ and applying control charts as described by Shewhart [9] and Pearson [5] to the observed y values, one may detect the first signs of a change in conditions "some time before this could be discovered by cruder methods, such as mere inspection of columns of figures."

7. Assaying an Unknown with the Aid of a Previously Established Relationship.

Having come this far, only one step farther is required to obtain a solution to a class of problems having the general nature of the following: A previously calculated regression, $\hat{Y}$, being available, a new value y' is observed and the value of X, say X', to which it corresponds has been lost sight of, or was never known. What value of X should be taken as the best single estimate of X', and within what limits can we assess X' with a confidence coefficient of .95 say?

From our previous discussion it is clear, I think, that in repeated sampling of *both* $\hat{Y}$ and y' the inequality (21) should hold 95% of the time, if $t_{.05}$ is the value for which $P\{|t|. > t_{.05}\} = .05$. Accordingly, *unless* our present experience with regard to $\hat{Y}$ *and* y' is in the upper or in the lower .025 tail of the t-distribution, y' is related to $\hat{Y}$ as indicated by (21). But the left side of (21) is really the same as the left side of (23) with $t_{.05}$ in place of $t'_{.05}$, and the right side of (21) can likewise be obtained from the left side of (23) by replacing $t'_{.05}$ by $-t_{.05}$, and in both cases y' corresponds to Q, X' being unknown as in the previous problem. In short, by setting $Q = y'$ in (24) and replacing $t'_{.05}$ by $t_{.05}$, we can use this revised (24) to obtain upper and lower limits for X', and *unless* our combined experience with regard to $\hat{Y}$ and y' is such as would occur 1 time in 20, the value of X which truly corresponds to y' will be within these limits. The "best" single estimate will be $X' = \bar{x} + \dfrac{y' - \bar{y}}{b}$, which can be obtained from (24) by setting $t = 0$, and it should be noted that the upper and lower limits of X' for a given confidence level are *not* symmetrical with respect to this value. With regard to the data of Figure 2, if our new value $y' = 45$, and if the confidence desired were merely .90 (so that we can use $t'_{.05} = t_{.10}$), the calculations yield $.7207 \leq X' \leq .8539$ with $X' = .7884$ as the best single estimate.

It is unlikely that a dairyman would ever be interested in obtaining limits for C/F from the F/DM value of a finished cheese, so that he would probably never have any use for this additional technique. On the other hand the preceding situation is a common one in connection with problems of biological assay where it is desired to evaluate the potency of a substance by comparing the response it produces, when administered to one or more animals, with a dosage-response relation previously established with dosages of known strength. In the preceding problem we considered the case in which y' was a single additional observation corresponding to an unknown X'. If, instead, we had $\bar{y}'$, the mean value of N' additional observations corresponding to an unknown X', it is clear that the denominator of (19) will be $\sqrt{S^2_{\hat{Y}'} + s^2_{y\cdot x}/N'}$ in this case, so that confidence limits for X' corresponding to a confidence coefficient of .95 will be

$$(25) \qquad X' = \bar{x} + \frac{b(\bar{y}' - \bar{y})}{C} \pm \frac{t_{.05}\cdot s_{y\cdot x}}{C}\sqrt{B(\bar{y}' - \bar{y})^2 + \left(\frac{N + N'}{NN'}\right)C}$$

and the "best" single estimate of X' will be

$$(26) \qquad X' = \bar{x} + \frac{\bar{y}' - \bar{y}}{b},$$

where $\bar{y}$ is the mean of the y's in the analysis of the original N values;
$\bar{y}'$ " " " " the additional N' y's corresponding to the unknown X';
b " " regression coefficient in (4');
$s_{y\cdot x}$ is given by (14) and depends on the scatter of the original y values about the regression Y, and is based on $N - 2$ degrees of freedom;
$B = 1/\Sigma(x - \bar{x})^2$, the summation being over the original X values;
$t_{.05}$ = two-sided 5% significance level of t for $N - 2$ degrees of freedom;
and $C = b^2 - t^2_{.05}\cdot s^2_{y\cdot x}\cdot B$.

In practice N' is usually small compared with N, so that $s^2_{y\cdot x}$ based on the original analysis will probably be used. However, if it is desired to make use of the dispersion of the new y' values to "improve" the estimate of $\sigma^2_{Y\cdot X}$, then $\bar{S}^2_{y\cdot x} = [(N-2)s^2_{y\cdot x} + (N'-1)s'^2]/(N+N'-3)$ should be used in place of $s^2_{y\cdot x}$, where $s'^2 = \Sigma(y' - \bar{y}')^2/(N'-1)$, and the $t_{.05}$ value corresponding to $(N + N' - 3)$ degrees of freedom used. Mathematically this is preferable to the above, but involves considerably more calculating, and probably would not be used by the practical man.

We shall illustrate the use of (25) and (26) in connection with the data of Figure 3 obtained from autopsies of 69 rats which had received doses of estradiol varying from 0.025 micrograms to .2 micrograms.[10] It was found that a linear relation, with a common variance on the various dosages, existed between $X = \log_{10}$ dose and $Y = \sqrt{\text{uterine wt.}}$ These are the quantities portrayed in Figure 3. The least squares line is

$$(27) \qquad \hat{Y} = 6.9023 + 3.4004(X + 1.0777) = 10.567 + 3.400X,$$

and is seen to be a good fit.

Carrying through the necessary calculations we find that 95% confidence limits for X', the true log dose, corresponding to a mean response of $\bar{y}'$ based on N' values, are

$$(28) \qquad \begin{aligned} X' &= -1.0777 + 0.2964(\bar{y}' - 6.9023) \\ &\qquad \pm .07074 \sqrt{.1376(\bar{y}' - 6.9023)^2 + \left(\frac{69 + N'}{69 \times N'}\right)(.09062)} \end{aligned}$$

and the optimum single estimate is

$$(29) \qquad X' = 0.2941\,\bar{y}' - 3.1077.$$

Dr. C. I. Bliss has informed me in correspondence that seldom is the sensitivity of an animal species to a hormone or other drug constant enough for the actual procedure outlined above to be reliable, so that in assaying any given sample it should always be tested in parallel with a standard preparation. If the slope of the regression, i.e. b, is fairly stable, even though the position changes, it is possible to assay the relative strength of an unknown by administering it and a standard at a single dilution, but it is preferable to use at least two dilutions in each assay so that it may be discovered whether the new b agrees substantially with that given by the standard dosage-response curve. Discussions of the procedures to be used in these cases will be found in references [26] to [31] from which I have received much inspiration.

[10] These data have been discussed by Lauson, Heller, Golden, and Sevringhaus [32] of the Wisconsin General Hospital, to whom I extend thanks for permission to use them in the present paper. Only a portion of their data have been used as the linear relation discussed below failed outside of the dosage limits given above.

8. **Concluding Remarks.** The formulae and ideas presented in this paper have been drawn in the main from the articles and books listed at the end of this paper. By turning to these references the reader often will find a fuller account of methods and applications than has been given here. In many cases the reader will find that the author of one of the references has placed emphasis on getting the answer. In the present paper the emphasis has been on the ideas and assumptions involved, the aim being to promote understanding of the methods discussed. In particular, the following two points have been stressed here:

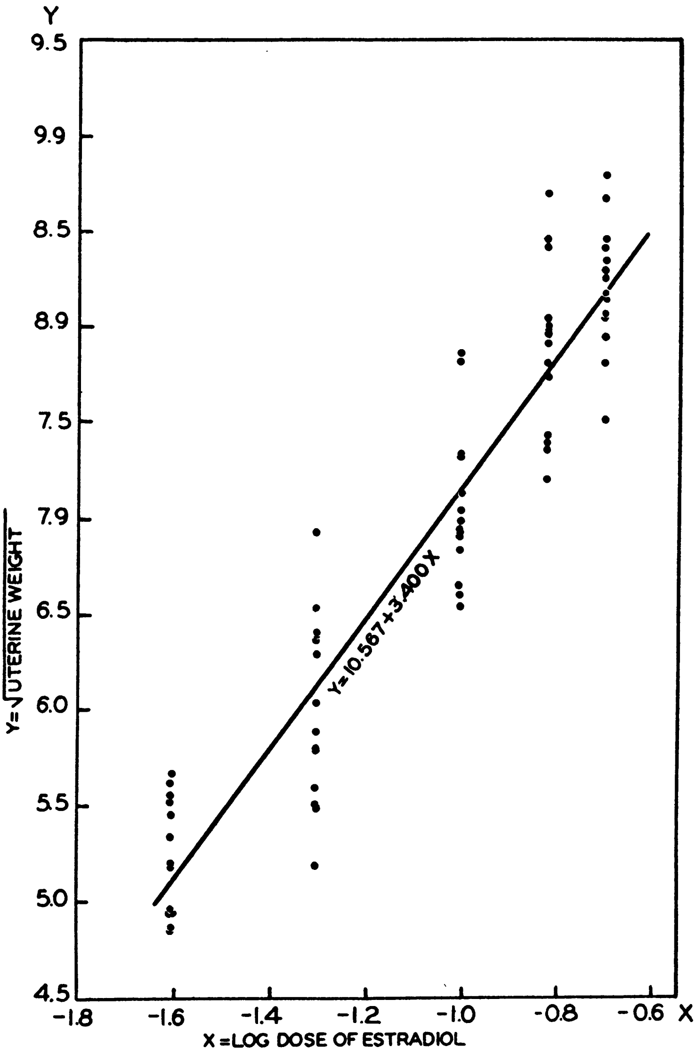

Fig. 3

(a) When the values of one of a pair of random variables are selected by the research worker, or when one of the variables is allowed to take values in only a restricted portion of its real range, then inferences with regard to an unknown value of this variable, say X, based on the corresponding (known) value of the other variable, say Y, are mathematically valid only when inferred from the relationship giving Y as a function of X; and

(b) The resulting inference is in the form of a confidence interval whose confidence coefficient is associated with the joint experience consisting of the observed regression of Y on X *and* the observed (or future) additional sample involving the unknown value of X, and *not* merely with the latter.

The ideas and assumptions which have been discussed have been illustrated on two examples.

Closer coöperation is possible between the practical man and the statistical theorist when the latter fully appreciates the problems of the former, *and* when the former, in turn, understands the methods advocated by the latter.

THE UNIVERSITY OF WISCONSIN.

REFERENCES

General Literature

[1] W. EDWARDS DEMING, *Some Notes on Least Squares.* Washington: Graduate School, Dept. of Agric., 1938.

[2] W. EDWARDS DEMING, "Some Thoughts on Curve Fitting and the Chi-Test," *Jour. American Statistical Association,* vol. 33 (1938), pp. 543–551.

[3] R. A. FISHER, *Statistical Methods for Research Workers.* Edinburgh: Oliver & Boyd, 6th edition, 1936.

[4] R. A. FISHER, "Applications of 'Student's' Distribution," *Metron,* vol. 5 (1925), pp. 90–104.

[5] E. S. PEARSON, *The Application of Statistical Methods to Industrial Standardization and Quality Control.* London: British Standards Institution, 1935.

[6] P. R. RIDER, "A Survey of the Theory of Small Samples," *Annals of Mathematics,* vol. 3 (1930), pp. 577–628. Copies available from Nordemann Publishing Co., N. Y. C.

[7] P. R. RIDER, *An Introduction to Modern Statistical Methods.* New York: John Wiley & Sons, 1939. pp. 78, 79, 90, 91.

[8] HENRY SCHULTZ, "The Standard Error of a Forecast from a Curve," *Jour. American Statistical Association,* vol. 25 (1930), pp. 139–185.

[9] W. A. SHEWHART, *Economic Control of Quality of Manufactured Product.* New York: D. Van Nostrand, 1931.

[10] W. A. SHEWHART, *Statistical Methods from the Viewpoint of Quality Control.* Washington: Graduate School, U. S. D. A. In press.

[11] "Student" (the late W. S. GOSSET), "The Probable Error of a Mean," *Biometrika,* vol. 6 (1908), pp. 1-25.

[12] HOLBROOK WORKING AND HAROLD HOTELLING, "Applications of the Theory of Error to the Interpretation of Trends," *Jour. American Statistical Association,* vol. 24 (1929), pp. 73-85.

Literature Relating to Fiducial Limits and Confidence Intervals

[13] R. A. FISHER, "Inverse Probability," *Proc. Cambridge Philosophical Society,* vol. 26 (1930), pp. 528-535.

[14] R. A. Fisher, "The Concepts of Inverse Probability and Fiducial Probability Referring to Unknown Parameters," *Proc. Royal Society of London,* Series A, vol. 139, pp. 343-348.

[15] R. A. Fisher, "The Fiducial Argument in Statistical Inference," *Annals of Eugenics,* vol. 6 (1935), pp. 391-398.

[16] J. Neyman, "On the Two Different Aspects of the Representative Method," *Jour. Royal Statistical Society,* vol. 97 (1934), pp. 558-625.

[17] J. Neyman, "On the Problem of Confidence Intervals," *Annals of Mathematical Statistics,* vol. 6 (1935), pp. 111-116.

[18] J. Neyman, "Outline of a Theory of Estimation Based on the Classical Theory of Probability," *Phil. Trans. Roy. Soc. London,* Series A, vol. 236 (1937), pp. 333-380.

[19] J. Neyman, *Lectures and Conferences on Mathematical Statistics*; edited by Deming; Washington: Graduate School, Dept. of Agric., 1938.

[20] J. Neyman, *L'estimation statistique traitée comme un problème classique de probabilité.* Actualités scientifiques et industrielles, Nr. 739 (1938), pp. 25-57. Paris: Hermann et Cie.

[21] H. L. Rietz, "On a Recent Advance in Statistical Inference," *Amer. Math. Monthly,* vol. 45, pp. 149-158.

[22] B. L. Welch, "On Confidence Limits and Sufficiency with Particular Reference to Parameters of Location," *Annals of Mathematical Statistics,* vol. 10 (1939), pp. 58-69.

[23] Albert Wertheimer, "A Note on Confidence Intervals and Inverse Probability," *Annals of Mathematical Statistics,* vol. 10 (1939), pp. 74-76.

[24] S. S. Wilks, "Fiducial Distributions in Fiducial Inference," *Annals of Mathematical Statistics,* vol. 9 (1938), pp. 272-280.

[25] E. B. Wilson, "Probable Inference, the Law of Succession, and Statistical Inference," *Jour. American Statistical Association,* vol. 22 (1937), pp. 209-212.

Literature Cited Relating to Methods of Biological Assay

[26] C. I. Bliss, "The Calculation of the Dosage-Mortality Curve," *Annals of Applied Biology,* vol. 22 (1935), pp. 134-167.

[27] C. I. Bliss, "The Comparison of Dosage-Mortality Data," *Annals of Applied Biology,* vol. 22 (1935), pp. 307-333.

[28] C. I. Bliss, "The Determination of the Dosage-Mortality Curve from Small Numbers," *Quarterly Journal of Pharmacy and Pharmacology,* vol. 11 (1938), pp. 192-216.

[29] J. H. Burn, *Biological Standardization.* Oxford Univ. Press, 1937.

[30] J. H. Gaddum, (British) Medical Research Council Special Report Ser. No. 183 (1933).

[31] J. O. Irwin, "Statistical Methods Applied to Biological Assays," *Supplement Jour. Royal Statistical Society,* vol. 4 (1937), pp. 1-48.

[32] H. D. Lauson, Carl G. Heller, June B. Golden and E. L. Sevringhaus, "The Immature Rat Uterus in the Assay of Estrogenic Substances, and a Comparison of Estradiol, Estrone and Estriol," *Endocrinology,* vol. 24 (1939), pp. 35-44.