STATISTICAL TECHNIQUES OF FORECASTING

By George W. Snedecor
Director, Statistical Laboratory
Iowa State College

Lecture Course in Highway Economics
Ames, Iowa
January 9, 1940

STATISTICAL TECHNIQUES OF FORECASTING

By George W. Snedecor
Director, Statistical Laboratory
Iowa State College
1940

A forecaster is one who, after observing a series of historical facts, undertakes to guess some related future event. He is not a prophet, but is a believer in trends. His materials are available data about the past. He may possess a single set of such data or groups of related sets. The essential thing is that his information must include the facts about some specified phenomenon in several past epochs.

From consideration of these facts, the forecaster must arrive at some opinion as to the future course of this phenomenon. He may do this by means of series of numbers, by means of graphs, by means of some elaborate methods of curve fitting, or through sitting in his chair and thinking out the answer. Some men really seem to know what is going to happen without aid of statistics. Most of us, however, deem it wise to record our data along with the methods used for reaching conclusions. This promotes objectivity and allows definite checks which, in turn, improve our skill. In this discussion I shall not consider the case of the armchair forecaster, but shall confine my attention to those who use statistical methods.

Fundamental in forecasting is the extrapolation of a series or a curve beyond the limits of the known data. Some such process is implied in nearly all good statistics. From a sample, my tribe of statisticians attempts estimates of the unknown parameters of the population from which the sample was drawn. As in your work, the sample itself often has little interest except in so far as it affords information about the population. The only difference between the problems with which you and I ordinarily deal is that my unknown population lies around me coexistent in time, whereas your unknown is in the future. With both of us it is the unknown which we attempt to fathom. Your techniques must include methods of extrapolation, and with these I shall deal below.

The character of the forecast required is usually twofold. First, there is the estimate of the event; and then, accompanying it, there are limits within which the event itself may be expected to lie. This latter introduces the probability theory of fiducial limits. Such duality in the nature of statistical thinking is deeply fundamental. Estimates and their probable errors are not two things which may either be taken or left, but are twin facades of the same structure. It may be that it is this feature of forecasting you wish me chiefly to discuss.

Forecasting implies faith in the continuity of nature. Now I don't know whether nature is continuous, or merely seems that way to the superficial observer. Perhaps what we call continuity is actually the statistical concept of average, a relatively stable aggregate made up of randomly varying elements. In your work you deal with social phenomena such as the invention of the automobile together with human propensities for mechanisms and speed. The associated trends apparently have been exhibiting a fair degree of continuity.

By training, the engineer is thoroughly inoculated with the stability of physical and chemical laws. He is prone to believe that concrete beams, fashioned in a specified routine, possess an immutable modulus of elasticity. True, his efforts to substantiate this seem to go wrong, yet he usually assumes that the average of a few variant determinations is likely to supplant the individual observation in endowment with the same illusive immutability. The concept and consequences of universal variation are more familiar to the biologist than to the engineer, but perhaps not more fraught with significance. Your request for a paper on the statistics of forecasting is perhaps an indication that you are becoming more acutely aware of the heavy role that variation plays in your experimental investigations, and are concerned with methods for taking account of it.

Nevertheless, despite variation, there must be some continuity even in the mileage of main roads. Deepseated human necessities, desires and emotions do

not change rapidly. Even revolutionary gadgets like the automobile must fit somehow into the social and economic patterns of life. If the series with which you deal indicate trends, there is warrant for using them in forecasting. The series used should furnish you not only an estimate for a future date, but also appropriate limits of error to safeguard the estimate.

The process of forecasting involves four procedures. First comes the collection and examination of series of data relating to the phenomenon to be predicted. While this is of fundamental importance, it is a procedure with which you are all familiar, and which I have not the information to discuss. Hence, I pass to the second procedure. Having decided on the date to be used, such functions must be selected and fitted as will lend themselves to extrapolation. These functions may be linear or curved. The fitting may be by rule of thumb or by some more objective scheme like the method of least squares. The third procedure follows immediately from the second, and may be considered part of it - the estimate of the desired quantity at a future time, together with the setting of probable limits within which the eventual value may be expected to lie. Finally, there is the interpretation of the statistical result, including the hypotheses upon which it is predicted, modifying circumstances which must be considered, and pertinent events whose occurrence would change the forecast. I shall write more at length of the second and third of these processes.

The selection of an appropriate function is the most difficult as well as the most critical part of forecasting. Many straight lines and curves will fit the given data almost equally well. How shall a choice be made among them? Which one affords the best indication of the future?

There is one consideration which I suspect is quite familiar to you. The curve which fits the data best may not furnish the best forecast. This is fundamental not only to the man who uses objective methods of fitting, such as least squares, but to the freehand fitter as well. Regular, sweeping curves

may leave the data points rather far afield, but they have greater validity on extrapolation. The person who subjects his curve to the erratic influence of every point has no course to steer by when the points are left behind. One reason for this is evident if the fitting of polynomials is considered. Suppose the series to be fitted consists of 10 annual values. A ninth degree polynomial can be made to pass through the ten points, but its behavior between them and beyond them is fantastic. A three, four, or five degree polynomial may behave nicely within the range of the data but is likely to rush off towards infinity immediately outside. This disadvantage of fitting a lot of constants is obvious to those who use objective methods, but the implications aren't always clear to the freehand boys. One risks some inaccuracy but gains emphasis by the generalization, "The fewer the constants the better the forecast."

The forecaster is sometimes fortunate in having a theoretical basis for choosing the function. His series may be increasing according to some growth law such as a geometric progression. While such instances are rare, any pertinent theory should be reflected in the equation which is adopted.

The selection of the function to be fitted is the severest test of the skill and foresight of the forecaster. Of two curves which fit equally well, he will choose that one whose behavior beyond the data seems most compatible with known facts. Functions with pronounced peculiarities of their own, such as the higher degree polynomials, may not follow the trend of the series at all beyond the confines of the known points. Such functions should be scrutinized with especial care.

Since the straight line is the simplest of all functions, it is ordinarily selected for forecasting unless there is adequate evidence in favor of some curve.

To illustrate the techniques of forecasting, I have selected the data in Table 1. For convenience, the dates have been "coded" by considering 1911 as year 1. The graph (Figure 1) indicates a remarkably uniform trend upwards,

Table 1. Speed Records Attained in Indianapolis Memorial Day Automobile Races

Year	X Coded date (Year - 1910)	Y Miles per hour
1911	1	74.7
1912	2	78.7
1913	3	75.9
1914	4	82.5
1915	5	89.8
1916	6	83.3
1919	9	88.1
1920	10	88.5
1921	11	89.6
1922	12	94.5
1923	13	91.0
1924	14	98.2
1925	15	101.1
1926	16	95.9
1927	17	97.5
1928	18	99.5
1929	19	97.6
1930	20	100.4
1931	21	96.6
1932	22	104.1
1933	23	104.1
1934	24	104.9
1935	25	106.2
1936	26	109.1
1937	27	113.6
1938	28	117.2
1939	29	115.0

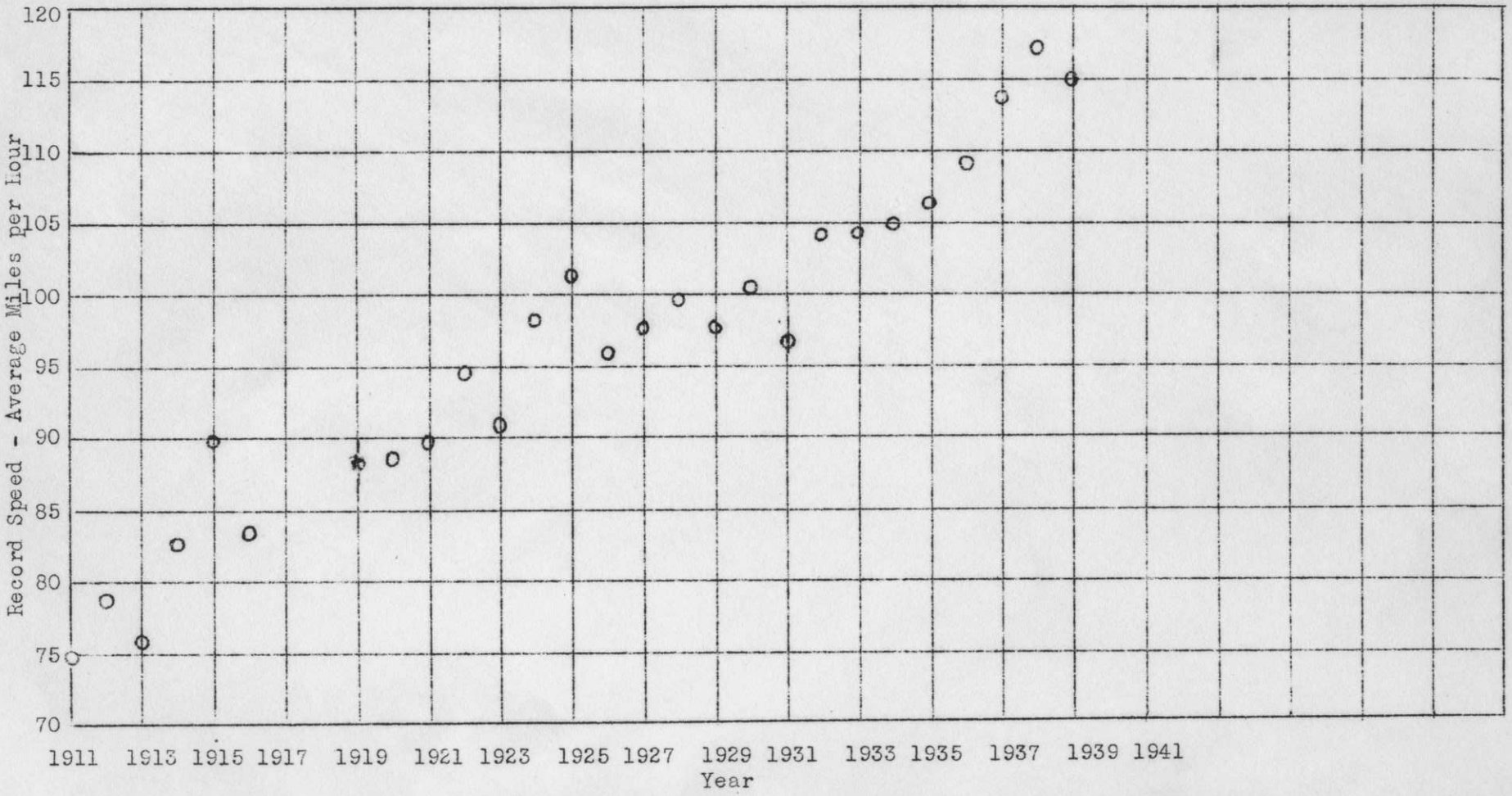

Figure 1. Automobile speed records, Indianapolis Memorial Day races.

though there is some indication of more rapid rises during the first and last six-year periods. This raises the question as to whether the real "population" trend is straight or curved. As a preliminary process, let us fit a straight line.

After the forecasting function is selected, the process of fitting it to the data is entirely objective and comparatively easy. Ordinarily the method of least squares is used. In applying this method, mathematical statisticians usually employ deviations from the arithmetic mean as convenient data. The formulas for the sums of squares and products of such deviations are

$$Sx^2 = SX^2 - (SX)^2/n, \quad Sy^2 = SY^2 - (SY)^2/n,$$

$$Sxy = SXY - (SX)(SY)/n,$$

the upper case letters being the observed dates (X) and corresponding speeds (Y) in the table, while n is the number of known speed records. Summation is indicated by S. The "corrections for mean," $(SX)^2/n$, etc., reduce the sums of squares and products of observed values, SX^2, etc., to sums of squares and products of deviations from means, the latter being denoted by the lower case letters. In what follows, the terms "sum of squares" and "sum of products," refer exclusively to these functions of deviations from means.

The straight line fitted by the method of least squares passes through $(\overline{x},\overline{y})$, where

$$\overline{x} = (SX)/n, \text{ and } \overline{y} = (SY)/n$$

are the means of X and Y respectively. The slope, b, is given by

$$b = (Sxy)/(Sx^2),$$

which, in the speed records, is

$$2,454.689/1,908.667 = 1.2861 \text{ miles per hour per year,}$$

the average annual increase in speed during the period covered.

The equation of the fitted line is

$$E = \bar{y} + b(X - \bar{x}) = 96.21 + 1.2861(X - 15.56)$$

$$= 1.2861X + 76.20$$

where E stands for the value of Y as estimated from the linear equation. This line is plotted in Figure 2.

For the forecaster interested only in an estimate, there remains merely the substitution of the desired year in the function with the resultant predictive value. But if you stop there, you have made inefficient use of the powerful mechanism which has been set up. Let me describe in some detail the further contribution which probability theory has to make.

We start by measuring the poorness of fit of the selected function, the deviations of the plotted points from the fitted line being pertinent data. You will readily see that the poorer the fit, the less the reliance which can be placed in the forecast. It turns out that the sum of the squares of the deviations (measured vertically) is useful information. This can be got, of course, by the tedious process of (i) substituting each X in the fitted equation and calculating the corresponding value of E, (ii) computing the deviation of each point, Y - E, from the fitted line, then (iii) squaring and adding these deviations, obtaining $S(Y - E)^2$. It is easily proved[1] that this result

[1]See, for example, Albert E. Waugh's "Elements of Statistical Method," McGraw-Hill Book Co., Inc., New York.

is quickly obtained by applying the formula,

$$S(Y - E)^2 = Sy^2 - (Sxy)^2/Sx^2$$

In our example, this gives

$$S(Y - E)^2 = 3,398.60 - (2,454.69)^2/(1,908.67) = 241.69$$

You will note the partition of the sum of squares, Sy^2, into two parts. The first, $(Sxy)^2/Sx^2$, is attributed to the hypothetical uniform change of speed through the years - it can be shown to be the sum of the squares of the deviations of the estimated values, E, from their mean, $\bar{y}$; that is,

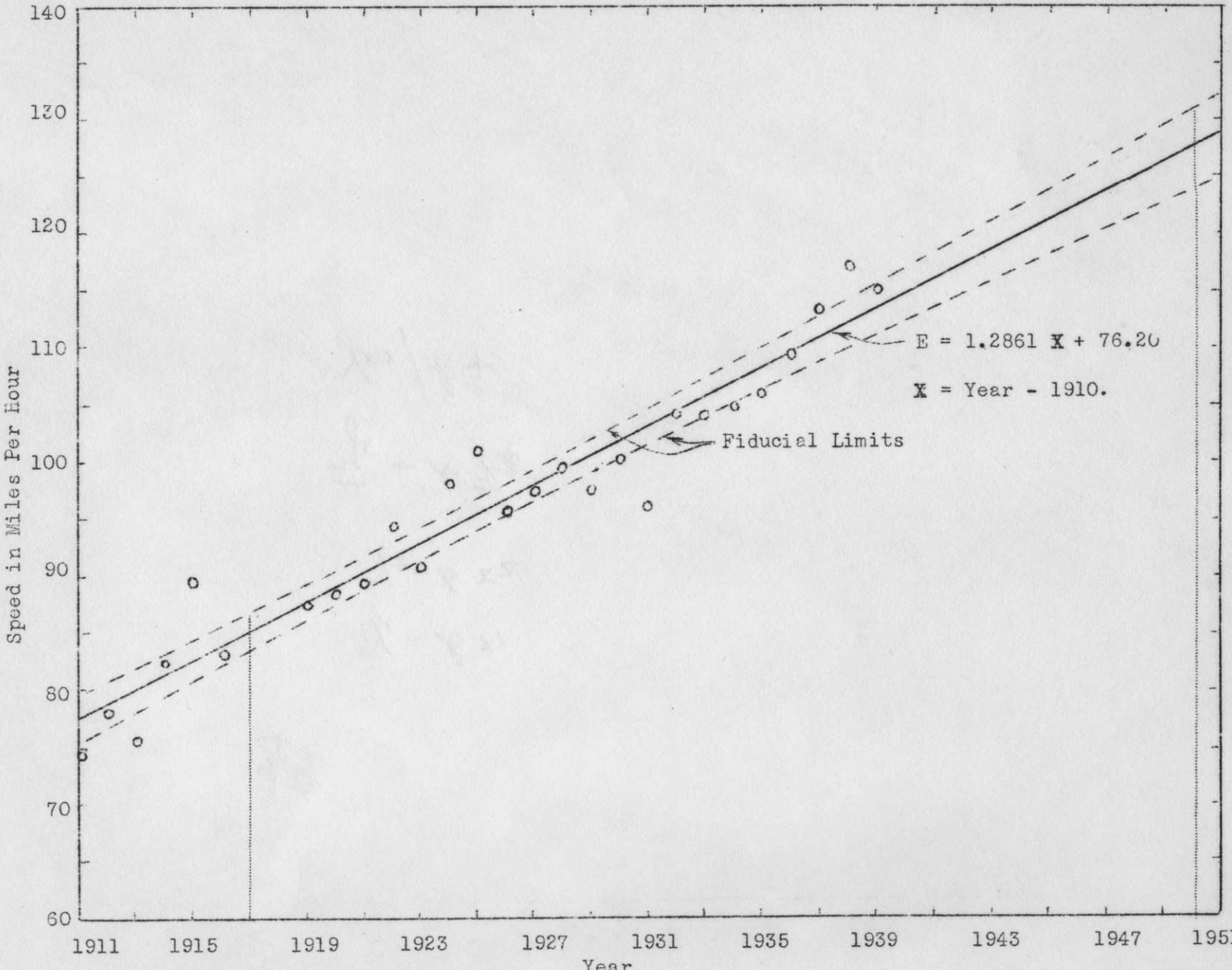

Figure 2. Straight line fitted to speeds by method of least squares, together with 95% fiducial limits.

$$S(E - \bar{y})^2 = (Sxy)^2/Sx^2$$

This would be identical with Sy^2 if each plotted point in Figure 1 had fallen exactly on the fitted straight line. The second part, $S(Y - E)^2$, being the sum of the squares of deviations from the fitted line, results from the failure of the plotted points to specify a uniform trend. This second portion deserves further comment.

The theory on which we are proceeding is that there is a "population" of speed records with linear trend, and that the deviations occurring in the annual races are due to a multiplicity of random causes having to do with weather, the drivers, accidents in the manufacture of tires, gasoline, etc. The random incidence of these "errors" is the basis of the probability theory involved. The set of 27 actual records in Table 1 is looked upon as a random sample of records, drawn from an assumed population which has linear trend. The two constants of our fitted straight line, $\bar{y}$ and b, are <u>estimates</u> of the corresponding "parameters" of the hypothetical population trend. If there were other samples, estimates from them would undoubtedly be different; that is, such estimates are subject to sampling variation assumed to be random. Finally, the random variation of $\bar{y}$ and b produces variation in E, also. It is this variation which is thought of as the source of poorness of fit, and which is associated with the second part of sum of squares, $S(Y - E)^2$.

Fisher[2] has devised a convenient scheme for summarizing the arithmetic above. Known as <u>analysis of variance</u>, it is applied to our speed record problem in Table 2. The term "degrees of freedom" characterizes the divisors of sums of squares appropriate for producing the mean squares needed for measuring variation. Each estimated constant is associated with one degree of freedom and with a portion of sum of squares computed in the manner heretofore described. This leaves n - 2 degrees of freedom for the remaining sum of squares, attributed to random error.

---

[2] R. A. Fisher. "Statistical Methods for Research Workers." Oliver and Boyd, Edinburgh.

Table 2. Analysis of Variance of Speed Record Data

Source of variation	Degrees of freedom	Sum of squares	Mean square
Total	27	$SY^2 = 253,306.96$	
Correction for $\bar{y}$	1	$(SY)^2/n = 249,908.36$	
Sum of squares	26	$Sy^2 = 3,398.60$	
Slope, b	1	$(Sxy)^2/Sx^2 = 3,156.91$	3,156.91
Remainder - random error	25	$S(Y - E)^2 = 241.69$	9.6676

We now have a measure of poorness of fit, called standard error of estimate,

$$s_{y.x} = \sqrt{S(Y - E)^2/(n - 2)} = \sqrt{241.69/25} = 3.1093$$

This is the standard deviation of the errors of estimate (or deviations from trend), calculated in the manner of Gauss.[3] (The number of observations, n, is often erroneously used as the divisor for estimating mean squares, instead of degrees of freedom, n - 1 and n - 2.) If the plotted points deviate greatly from the trend line, $s_{y.x}$ is large. If the points should all lie on the line, $s_{y.x}$ would be zero.

[3] Carl Friedrich Gauss. Theoria Combinationis Observatorium, Pars Posterior (1821), Werke 3:31.

The standard deviations of both $\bar{y}$ and b may now be computed. For simplicity, I shall use the square of the standard deviations, known as variance, V. From the foregoing paragraph,

$$V_{y.x} = 9.6676$$

We now have, from known formulas,

$$V_{\bar{y}} = V_{y.x}/n = 9.6676/27 = 0.3581$$

$$V_b = V_{y.x}/Sx^2 = 9.6676/1908.67 = 0.005065$$

The square roots are measures of the variation of the mean, $\bar{y}$, and of the slope of the trend line, b.

Let us revert to the matter of estimating the population speed at a given date by substituting a value of X in the trend equation. For example, suppose

we estimate the population speed in a year when there was no race, 1917 (X = 7):

$$E(1917) = 96.21 + 1.2861\ (7 - 15.56) = 85.20 \text{ miles per hour}$$

Such an estimate is subject to two sources of variation. It will be in error if the mean, $\bar{y}$, is incorrectly estimated, and also if the slope b is not the same as the population value. If $\bar{y}$ is in error, the fitted trend line is too high or too low, and E is subjected to this hazard equally in every year. But incorrect estimates of the slope b result in rotations of the line about its true position, and this hazard affects E more seriously as X moves away from the mean, $\bar{x}$. It has been shown[4] that the variance of an estimated value, E,

[4] Holbrook Working and Harold Hotelling. "Applications of the theory of error to the interpretation of trends." Journal of the American Statistical Association, Vol. 24. March 1929 Supplement.

is given by

$$V_E = V_{y.x}\ (1/n + x^2/Sx^2)$$

For 1917, $x = X - \bar{x} = 7 - 15.56 = -8.56$ years. Substituting,

$$V_{1917} = 9.6676\ [1/27 + (8.56)^2/(1908.67)]$$

$$= (9.6676)(0.07543) = 0.7292,$$

the corresponding standard deviation being $s_{1917} = \sqrt{0.7292} = 0.854$ miles per hour. This is the measure of the poorness of estimate that must be faced at a time 8.56 years removed from the mean time, $\bar{x} = 15.56$ years (measured, you remember, from 1910).

From tables of the distribution of a function called t[5] we may select a

[5] "Student." "New tables for testing the significance of observations." Metron 5, No. 3(1925).

certain value corresponding to available degrees of freedom and to specified levels of probability of occurrence, and on either side of the estimated value we may then lay out upper and lower probable limits to the true value of the

speed record. Fisher[6] has designated these as fiducial limits. For 1917

[6] R. A. Fisher. Proceedings of the Cambridge Philosophical Society, Vol. 26, p. 528 (1938).

estimated speed, the fiducial limits, corresponding to 25 degrees of freedom and to t = 2.060 (the 5% level of probability) are

$$E_{1917} \pm (t_{.05})(s_{1917}) = 85.20 \pm 1.76,$$

that is, from 83.44 to 86.96 miles per hour. Concerning these limits we may make either of these statements, the second being explanatory of the first. (i) The fiducial probability is 95% that the population speed for 1917 lies between 83.44 and 86.96 miles per hour. (ii) If we repeat over and over this whole process of estimation from samples drawn at random from a population with linear trend, and if each time the statement is made that the true value lies within the fiducial limits calculated from the sample in hand, then 95% of such statements will be right.

The locus of the function,

$$\bar{y} + bx \pm t_{.05}\sqrt{V_{y.x}\,(1/n + x^2/Sx^2)}$$

is a pair of hyperbolas lying on either side of the trend line. These are plotted in Figure 2, in which the 1917 fiducial limits, calculated above, can be read off. Near the middle of the known time series, where the deviation x is small, the first term under the radical, involving the sampling variation of the mean, is usually larger than the second. But in forecasting, the second term, involving the variation of b, is usually predominant.

It must be clearly understood that we are attempting to estimate the population speed at different dates, assuming linear trend. If it is desired to set fiducial limits on individual speed records, allowance must be made for the sampling variation of single items with mean estimated as above. A method has been proposed[7], but will not be considered here.

[7] C. Eisenhart. "The interpretation of certain regression methods and their use in biological and industrial research." The Annals of Mathematical Statistics, Vol. 10, p. 162 (1939).

Let us proceed to forecast the speed for 1950. We shall be assuming that the population trend is really a straight line, and that our only hazard is the random variation inherent in the estimates of $\bar{y}$ and b. Substituting in the forecasting function:

$$X = 1950 - 1910 = 40,$$

$$x = 40 - 15.56 = 24.44$$

$$t_{.05} = 2.060$$

we have,

$$E_{1950} = 96.21 + 1.2861\ (24.44) = 127.64$$

$$V_{1950} = 9.6676\left[(1/27) + (24.44)^2/(1{,}908.67)\right]$$

$$= (9.6676)(.34999) = 3.3836$$

$$s_{1950} = \sqrt{3.3836} = 1.839$$

$$E_{1950} \pm s_{1950} = 127.64 \pm 3.79,$$

that is, from 123.85 to 131.43. Continuing the foregoing substitutions for other years, we get the graph of Figure 2.

In selecting 95% as the basis of the fiducial limits, I was merely following current practice in the field of biological experimentation. You may, of course, set corresponding limits for any desired probability.

I think every forecaster, whether he uses objective methods or not, must in effect go through some such mental process as the setting of fiducial limits. He knows that his estimate cannot be exact, and he has some idea of the variation to be expected. The advantage of using standard statistical methods is that the mental peculiarities of the individual forecaster are nullified - the limits as well as the estimate are numerical, and would be identically arrived at by every person choosing the same function.

The question may now be raised whether the linear trend we have been using is justified. Since the records increased more rapidly during the early and late years, a cubic might be more appropriate. Probability theory gives some

information as to goodness of fit. If we fit the cubic,

$$E = a + bX + cX^2 + dX^3,$$

we may then analyze the variance as follows[8]:

[8]For statistical methods used, see George W. Snedecor, "Statistical Methods." Collegiate Press, Inc., Ames, Iowa.

Table 3. Analysis of Variance of Speed Records Fitted With 1st and 3rd Degree Polynomials

Source of variation	Degrees of freedom	Sum of squares	Mean square
Remainder in Table 2	25	241.69	
Remainder if cubic is fitted	23	173,89	7.56
Increased precision of 3 constants over 1	2	67.80	33.90
F = 33.9/7.56 = 4.48*			

The meaning of all this is that the chances are less than 5 in 100 that so large a sum of squares would be accounted for by two extra constants if the population regression were really linear. From the character of the variation in the sample, one may well question the hypothesis of linear trend in the population.

But, having considered the statistical evidence, I must sit in my armchair and think. The cubic curve is turning up at the right with increasing acceleration. Shall I commit myself to the proposition that speed records are going to increase more and more rapidly from year to year? With the limited knowledge I have, I reject the idea. Of course, a new speedway may be built, new cylinder capacities may be allowed, and other changes in the rules may be made. But such would constitute a new population to which my estimates have no application. Forecasting new populations is not our problem. So far as my evidence goes, I would trust the linear hypothesis rather than the cubic.

Despite the evidence of the sample, I should be inclined to seek a function with a negative acceleration, one perhaps approaching a horizontal asymptote.

Such functions are available[9], but will not be discussed here.

[9]Henry Schultz. "The standard error of a forecast from a curve." Journal of the American Statistical Association, Vol. 25, page 139 (1930).

To this point I have considered in detail the use of only a single series. Perhaps you will be faced with several related series which must be combined into a single forecast. Take traffic density, for example, on the Ames-Des Moines highway. If you know this density at given times in the past, you may be satisfied to forecast from the single set of data. You will be assuming that the elements entering into the trend will continue to influence it in the future as in the past, and that each of these elements will maintain its trend. There may be cases in which these assumptions are not valid. Some elements may cease to function and others may enter the picture. In statistics we think of the methods of multiple regression as suitable when one variable is to be estimated from several others. Multiple regression is merely an application of the least squares method of fitting where two or more independent variables are involved. It is subject to treatment based on probability theory in the same way as described above for simple regression. Within its field of applicability, it is a well established statistical method, easily applied. I am unable to say whether or not it may be applied to the problems arising in your investigations. If not, then your problems may have to be tackled independently by some related scheme. Theoretically you will always be led to the twin solution - an estimate and a pair of fiducial limits. The derivation of these quantities is the problem discussed in the statistical "Theory of Estimation."

Finally, I should like to reemphasize a point which has been made over and over and over again in the foregoing remarks. You, of course, have it constantly in mind, but it is likely to be neglected by wild-eyed enthusiasts as well as by ignorant persons who have been heavily inoculated with statistics. It is easy to talk about and easy to calculate least squares estimates,

fiducial limits, etc., but it is only the wise who can be trusted to draw conclusions from the results. Statistics is a tool useful for objective evaluation. When it is looked upon as an end in itself, or when it is applied blindly, it may serve to obscure rather than to clarify. If I were forced to choose between the predictions of an armchair forecaster with wisdom and a highpowered statistician who sees no further than his calculating machine, I should take the armchair every time. My ideal is to provide the wise man in the chair with the services of the statistician and his machine.