THE RÔLE OF STATISTICAL METHOD IN ECONOMIC STANDARDIZATION

By W. A. SHEWHART

Abstract of lecture presented at the University of London, May, 1932.

1. OBJECTIVE

IMAGINE living in a world without standards—no signs, no words, no symbols for things which *mean* the same to all. Imagine carrying on a business like that of the Bell System without standards. Take, for example, the telephone itself: it is not so simple as it looks. To make it requires 201 parts; and to connect it with another instrument requires approximately 110,000 other parts. The annual production of most of these parts runs into the millions, so that the total annual production of parts runs into the billions. The manufacturing organization, the Western Electric Company, buys for use in its plants more than 25,000 different items of materials and supplies, and these come from more than 15,000 suppliers located in every state in the Union.

Let us examine some of the fundamental problems involved in establishing standards of quality—standards by which the consumer may judge the quality of product, and which in themselves represent the goal of the producer—and in so doing point out the rôle of statistical methodology. We shall consider some of the problems involved in establishing what we shall term *economic standards*—those where, under the given conditions in respect to the development of science and the development of human wants, there is a balance between the economic value to the consumer of any possible modification in the quality standard and the cost of such modification. To begin, let us consider the question:

2. WHAT IS A STANDARD?

I have before me as I write this paragraph a book of several hundred pages published by the American Society for Testing Materials. Its title is *A.S.T.M. Standards, Part I, Metals, 1930*. As is true of almost every other book, the only things I find inside are some symbols, lines, and pictures. These are supposed to specify some of the things which consumers want and producers are to make. It is essential therefore that they *mean* the same to both.

But when do symbols mean the same to a given group of people? The modern logician will likely answer that a symbolic statement means the same to two or more persons when it leads to the *same action* on the part of each of them.

If I were to ask the question:

$$2 + 2 = ?,$$

all of us would answer 4. If an officer of the law were to enter and say—"Leave this room immediately"—there would be a likeness in our actions. If, however, each of us were to try to make on a piece of paper ten straight lines of the same length, it is almost certain that the lines would not be identical. Now, go one step further, and imagine specifying such a simple thing as a wire nail so that the action on the part of every one of several producers would be the same.

There are at least two ways used by engineers to determine when two or more things are the same. One is for those interested to agree that they are the same for all intents and purposes. The other is for those interested to agree to a method of measuring the things, thereby getting as many sets of numbers as there are things, and further, to agree to a method of comparing the sets of numbers so obtained. Note that in either case the determination is simply an agreement. To settle upon a method of measurement again introduces the difficulty of determining when the method agreed upon—necessarily symbolic in form—means the same to all.

Thus the specification of a standard is of value only in so far as it indicates a way of arriving at an empirical decision acceptable to both producer and consumer as to whether or not a given thing is to be considered as being of standard quality. Only through agreement either upon the human judge or upon the formal processes of analysis of quantitative data can we attain this acceptable decision.

As a specific illustration, let us consider the standard of length—perhaps the simplest in existence. There is the concept, corresponding to the essence of the thing itself and represented customarily by the symbol L. Associated with this, there is that of the quantity of length involving a conceptual method of measurement. In other words, we think of so much of something—in this case so much of that which we designate L.

Suppose we have a number n of things whose lengths are to be measured such, for example, as straight lines. First we conceive of these lines having lengths measurable by some ideal process such that the degrees of so-muchness of the lengths are in point of principle expressible by numbers

$$X_1', X_2', \cdots, X_i', \cdots, X_n', \tag{1}$$

in terms of the chosen unit of length.

The actual process of measuring these same lines gives us, however, a corresponding set of numbers,

$$X_1, X_2, \cdots, X_i, \cdots, X_n. \quad (2)$$

Fundamentally (1) is conceived to exist, but we do not know the numbers; (2) is known and represents the results of applying a given rule of procedure of practical measurement for establishing such a set of numbers. Set (2) is that which we use as a basis for inferring something about the unknown set (1).

Suppose that we take the length of one of the lines as the standard of comparison and ask if the others are to be considered as being of this standard length. The only thing that we can do is to try to answer this question from a consideration of the set of numbers (2), no one of which may represent the true length of the corresponding line. Obviously to get any place, we must agree to a method of interpreting the data. This example shows that we cannot expect the specification of a standard to outline a way of determining whether or not a thing *is* of standard quality. This is true even when what we mean by quality is limited to a few measurable characteristics, for we always have to allow for errors of observation. To the question, *Is* a given thing of standard quality? there can be no positive answer, yes or no. This is simply another way of saying that all inference is of the nature of probability or degree of belief inference.

This situation is of far reaching importance in the problem of standardization. There is always a certain amount of variability which must be left to chance. The economics of the situation gives a basis for arriving at a definite answer to the question—How much variability shall be left to chance?

The postulational basis for answering this question is: A variation should be left to chance only when produced by a constant system of chance or unknown causes among which no single cause or distinguishable group of causes appears to have a predominating effect. This means, of course, that the practical test for constancy of this type is an agreed-upon criterion, or set of criteria, based upon *statistical* theory and particularly on the notion of statistical stability.

3. The Notion of Statistical Stability

Perhaps the simplest observable type of statistical stability is that defined as follows: Whenever an event may happen in only one of two ways, and the event is observed to happen under the same essential conditions for a large number of times, the ratio p of the number of times the event happens in one way to the total number of trials appears to approach a definite limit, let us say p', as the number n of trials increases indefinitely. Here p' is taken to be the statistical probability. Symbolically we may state this Law of Large Numbers in the form

$$\underset{n \to \infty}{L_s} p = p', \tag{3}$$

where L_s stands for the statistical limit which differs from a mathematical one in that we do *not* reach a number n_0 of trials such that, for all values of n greater than n_0, the ratio p will differ from some fixed value p' by less than some fixed positive value ϵ.

However, this formal definition is meaningless until we go further and consider how in practice we may put meaning into the phrases "an essential condition," and "statistical limit." We shall try to do this by a series of examples.

A typical case in point is that of tossing a coin and observing the successive values of the ratio p of the number of heads to the total number n of throws as the number n is increased indefinitely, as is illustrated by the results of one thousand throws shown in Fig. 1. I

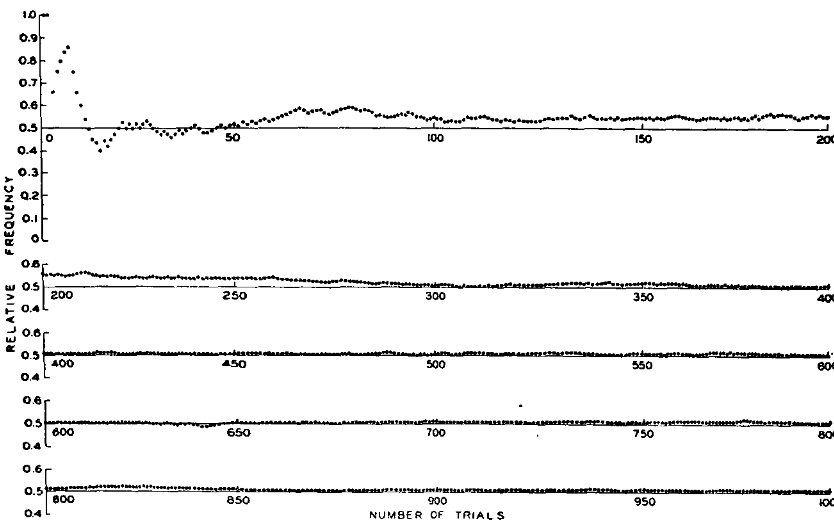

FIGURE 1. AN EXAMPLE OF STATISTICAL CONSTANCY.

have discussed in some detail elsewhere[1] several of the important characteristics of this limit of interest to engineers. For our present purpose, it is sufficient to call attention to the fact that formal mathematical criteria exist for assisting in determining when a set of data gives evidence of this kind of statistical constancy, and that these criteria have been satisfied in cases coming to our attention where those interested have been in a place to agree that the same essential conditions had been maintained.

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

As another example, let us consider the objective charge X' on an electron. It is generally assumed that this is a constant. A series of measurements of this charge, however, do not show constancy when taken even by an outstanding physicist like Millikan. In fact our only approach to the assumed objective reality is statistical. All we can say is that the average of a number n of measurements appears to approach a statistical limit. Actually, successive averages of 1, 2, 3, . . . of Millikan's measurements form a sequence of much the same kind as the one pictured in Fig. 1.

But in the present era of physical science there are very few instances where we even assume, as in the electron example, an objective constancy on the part of measurable characteristics. Instead we assume that the objective thing measured is itself statistical in nature. Witness, for example, any physical property such as density. Fig. 2 shows the

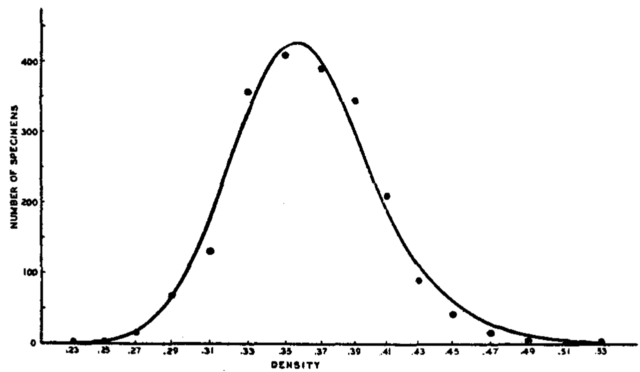

FIGURE 2. EXAMPLE OF STATISTICAL NATURE OF PHYSICAL PROPERTIES.

observed distribution of density determined on 2,105 small clear specimens of Sitka spruce, the wood at one time used extensively in making airplane propellers.[2]

In this case, care had been taken to eliminate causes of variability by careful selection of experienced engineers. Yet we note the wide variability in the measurements. Of course, this is partly attributable to errors of observation but, in this case, the error effect is small. Our best guess as to the character of the objective frequency function is represented by the solid curve.

As yet another example, consider a *Functional or Mathematical Law or Relationship:* One of the best known specifications of this type is

[2] L. J. Markwardt, "Comparative Strength Properties of Woods Grown in the United States," *Technical Bulletin No. 158, U. S. Dept. of Agriculture,* February, 1930.

the equation of state of a perfect gas, $PV = RT$, where P is the pressure; V, the volume; T, the temperature; and R the gas constant.

Taking the simplest case of two variables Y' and X', let us represent the assumed objective constancy by the functional relationship,

$$f'(Y', X, \alpha_1', \alpha_2', \cdots, \alpha_i', \cdots, \alpha_m') = 0,$$

where the α''s are unknown parameters.

In practice what we really observe is that whenever our measurement X of the independent value X' takes on a particular value, let us say X_1, the ratio p of the number of times that the corresponding value Y will be found within any limits Y_{11} to Y_{12} to the total number of trials satisfies the statistical limit (3).

Similarly we may consider a *Statistical Relationship:* Two variables are said to be statistically related under the conditions:

(a) That the observed fraction p of the number of times that simultaneously observed pairs of values of these two variables fall within a given area in the plane representing these two variables, approaches as a statistical limit L_s some fixed value p', as the number n of observations approaches infinity, or in other words that we have (3).

(b) That the frequency distributions in the array of one variable are not the same for all of the arrays.

The great majority of relationships used as a basis for specifications are perhaps *Empirical Relationships.* This type of relationship may be exemplified by the assumed objective relationship between two physical properties of a metal. What we said about the statistical limit in our discussion of the functional relationship applies here. The only difference between this kind of empirical relationship—as we shall consider it here—and the functional relationship is that it is generally too complicated to express in a practical functional form.

We may conclude that: *Although we may conceive of various kinds of constancy as a basis for specifications of standards, the only kind of observable constancy is statistical.*

It is obvious, therefore, that the statistical nature of observable data must be taken into account both from a design and an inspection viewpoint. That is to say, the design engineer must be given information that will permit him to allow in a rational way for variability which must be left to chance. The inspection engineer needs to use statistical criteria to determine when the variability is greater than it is economical to leave to chance.

4. The Statistical Aspects of Human Wants

Let us ask a very simple question: What are the economic standard lengths of wire nails? It is conceivable that nails of every length between zero and some fixed upper limit have economic value: but needless to say, it is not feasible to manufacture nails of every length within this range.

History shows that the present commercial sizes of nails just came into existence—*laissez faire*, the economist would say. In other words, the present choice of sizes represents an agreement between consuming and producing agencies. It is sometimes argued that the practice of making nails of such sizes is of so long standing and has led to so many associated practices that it is quite reasonable to believe that these sizes may be accepted as economic standards. Is this conclusion justified? Possibly it is in this and some similar instances. Nevertheless it is not difficult to point to standards which came into existence in somewhat the same way and which later were found to be uneconomical. Witness, for example, the systems of weights and measures, many of which cannot be said to have been chosen with due consideration to the economic consequences of the choice.

One thing is certain, engineers today are in many instances giving very careful consideration to the choice of sizes when establishing new standards as, for example, in the assignment of wave lengths to broadcasting stations. The history of such assignments reveals less of the method of *laissez faire* and more of national and international planning in which an attempt has been made to weigh carefully the economic consequences of the assignments. In fact, we have today many national and international standardization organizations giving careful consideration to choice of sizes of quality characteristics of one kind or another.

Looked at closely, we see that in all cases the choice of economic standard sizes involves the problem of measuring the relation of *consumer demand* to the choice of the system of sizes. Needless to say, this involves a sampling problem of basic importance, and it is reasonable to believe that few systems of sizes existing today would bear close scrutiny from this viewpoint. In this we have a fundamental economic problem, statistical in nature, which must be given careful consideration as a part of any attempt in either national or international planning having to do with the establishment of sizes or aimed-at values of quality characteristics that will give the greatest feasible satisfaction to the maximum number of people.

My attention was first drawn to the sampling problem involved in determining human wants in connection with the specification of sizes of sound-proof aviation helmets. A similar problem is the determination of the best dimensions of a telephone handset.

In determining sizes and allowable tolerances, one of the statistical factors of interest is the distribution of the minimum detectable increment ΔX in a given characteristic for a homogeneous group of people.[3] If we vary two or more characteristics detectable by the senses, such as frequency and intensity, the objective distribution is one in two or more dimensions.

Most frequently sensation measures are of this complex sort. Take for example the noise current of a telephone circuit. I have discussed elsewhere[4] how the control of such a phenomenon requires the calibration of a machine method of measurement of the current which must be correlated to the ear measure of a homogeneous group.

The same kinds of problems are involved in color testing and in any other case where a direct sensation is at the basis of our judgment of whether or not we like or want a thing.

Needless to say, how much we want a thing depends upon what we think we know about it, and about potential things of its kind that might be made. A Robinson Crusoe would no doubt be happy with an old-fashioned radio, but his relatives must have the latest model; and the radio engineer must have something even better in his home.

But what makes each one want what he wants? There's the rub. *To determine the causal effect of a measurable characteristic of reality on our want for a thing is most assuredly a sampling problem in which experiments must be carefully planned to get the correlation between corresponding variates as high as possible so as to reduce the sampling error to a minimum.*

5. Efficient Production Method

Assuming that the end requirements can be specified in a manner agreed upon by producer and consumer, let us next consider the statistical aspects of establishing an efficient production method.

Some of the problems[5] involved are: The determination of ways and means of effecting economies through plant location, design, choice of material, purchasing, and marketing; the economic balancing of labor and machines; economic consideration in selection and purchase of equipment; efficient use of capital investment in equipment; and economic materials handling, to mention only a few.

If we examine almost any one of these problems in detail, it is but natural that we should find that the sought-for solution depends upon both physical and economic factors. The industrialist charged with the responsibility of making scientific developments of greatest use to the maximum number of people finds that the equations involving the cost

[3] W. A. Shewhart, "Some Applications of Statistical Methods," *Bell System Technical Journal*, January, 1924.

[4] *Economic Control of Quality of Manufactured Product, loc. cit.*

[5] Cf. Eidmann's *Economic Control of Engineering and Manufacturing*, McGraw Hill Publishing Company, 1931.

factor which he wishes to minimize usually, if not always, contain these two kinds of elements.

Two Bell Telephone System colleagues, S. L. Andrew[6] and R. W. Burgess,[7] both actively engaged in the application of statistical methodology to the economic aspect of these problems, have recently discussed better than I can the statistical nature of the problems involved and the need for better data—data that have a common and well defined meaning—to be used in the available statistical machinery.

What is needed most perhaps is a changed viewpoint in the application of statistics to these economic problems—*away from the summarizing objective in the form of indexes whose physical meaning is indeterminate and toward the separation of data into rational subgroups with the assistance of the latest improvements in statistical criteria.*

Turning next to the physical side of the problem, we find one of the important fields of application of statistical theory in the establishment of distribution of tolerances on piece parts or quality characteristics to come within the overall allowable tolerances. There are two types of design from this viewpoint: one in which the overall quality is a function of the qualities of component parts in such a way that the resultant deviation is of an additive kind; the other in which the resultant deviation is of a chain type in respect to the component parts.

Let us consider a very simple problem of the first type. Suppose that you are building a rack to support a load consisting of the combined weights of a number of different pieces of apparatus. Assume that past experience is available to estimate the average and the standard deviation in the weights of each of the kinds of apparatus. One very customary method in such instances is to allow for the maximum load plus a certain safety factor where the maximum load is taken as a sum of the maximum weights that have ever been observed. Of course the chance of obtaining this maximum load is negligibly small, and there is little engineering justification for designing for such a condition because the assurance attained in this instance is out of all proportion usually to the assurance attained at other points in the system. To take a simple case where the standard deviations in all the weights are equal, the satisfactory maximum load is such that the amount added to allow for dispersions in the separate weights need be only $1/\sqrt{n}$ times as large as that given by the customary method. In other words, the

[6] "The Methods of Industrial and Business Forecasting," S. L. Andrew, Chief Statistician of the American Telephone and Telegraph Company, *Bell Telephone Quarterly*, January, 1931.

[7] "A Statistical Approach to Mathematical Formulation of Demand-Supply-Price Relationship," R. W. Burgess, Chief Statistician of the Western Electric Company, *The Annals of Mathematical Statistics*, February, 1932.

method often used gives $\sqrt{n}$ times the necessary addition in strength with its associated cost.

Now, let us consider the second type of problem, or that of specifying the distribution of the breaking strength of a chain composed of m links, the breaking strengths of the links used in the chain being assumed to be normally distributed about a specified value $\overline{X}'$ with a standard deviation σ'. Obviously the more links there are in the chain, the lower becomes the expected breaking strength of the chain. The objective standard deviation also decreases. The work of Tippett[8] makes it possible for the design engineer to estimate this effect accurately.

To make these applications possible in a given design it is of course necessary that the qualities of separate piece parts be *statistically controlled* in the process of production. R. L. Jones[9] was one of the first to grasp the significance of the condition of statistical control as a basis for inspection engineering, not only because it leads to maximum consumer protection, but also because it is the logical basis for inspection procedures to attain the following five objectives: (a) To determine for each step the economic percentage rejection, or tolerance p' for defectives; (b) To reject defective material and parts at such points in the chain of production as will make the net cost of rejection a minimum; (c) To determine for each step the minimum amount of inspection which will suffice to give economic control of quality; (d) To detect failures of desired quality control as evidenced by fluctuations and trends; (e) To discover the causes of such fluctuations and trends in order to secure improved control.

The first object is, as it were, to insure that the quality of material passed from one stage to the next is as good as we ought to make it, all the costs both before and after this stage being taken into account. The tolerance for defects under these conditions will depend on at least three essential factors: (1) the cost of refining the earlier processes, as by improved machinery or more skillful personnel, (2) the cost of inspection to eliminate difficulties at a given stage, and (3) the cost of remedying defects at a subsequent stage. Stated in another way, the requirement amounts to saying that we should not let the observed fraction p defective fluctuate more than it is economical to leave to chance. But this is the condition of statistical control for economy of

[8] "On the Extreme Individuals and the Range of Samples Taken from a Normal Population," *Biometrika*, Vol. XVII, December, 1925, pp. 364–388. These tables are also reproduced in the second volume of Pearson's *Tables for Statisticians and Biometricians*, 1932.

[9] Director of Apparatus Development of the Bell Telephone Laboratories. "Quality of Telephone Materials." *Bell Telephone Quarterly*, Vol. VI, pp. 32–46.

production. Moreover the observed fraction defective at a given stage does not give a very good picture of the portion of the material that is not sampled unless assignable causes have been weeded out in previous stages. Thus from all angles it appears that the first step essentially requires that the product be statistically controlled in order to attain the first objective.

If a quality of a product is controlled in this sense the probabilities of getting $0, 1, 2, \cdots, n$ non-conforming pieces in a lot of n pieces are obviously given by the terms of the point binomial $(q'+p')^n$. The practical problem of inspection is therefore that of watching the observed fractions $p_1, p_2, \cdots, p_m$ say, as determined from samples of $n_1, n_2, \cdots, n_m$, to see if they give any indication of the existence of assignable causes of variation. From this viewpoint the emphasis in the inspection specification should be on the collection of data separated so as to detect assignable causes, if they exist, more than on the inspection of a fixed portion of the product as is often done. The best results will naturally be obtained by the engineer who is most successful in dividing his data into objective rational subgroups.

As soon as the assignable causes have been detected and removed one after another, we may expect oscillations of the fraction p similar to those in the observed ratio of heads to total number of throws of a coin. In other words, starting with the observed fraction in a sample of a single piece of product it will be either 0 or 1 and will then oscillate about some fixed value much in the same manner as the frequency in Fig. 1.

For reasons which we have already considered, so soon as we have secured statistical control we have maximized the amount of assurance given by a sample; and we are in a position to say in most instances that the rate of increase in the precision of our results is roughly proportional to the square root of the number of observations. Thus objective (c) is attained. Obviously, in attaining the state of statistical control, we have attained objectives (d) and (e).

An example will illustrate the need for statistical analysis of *variability* in problems of the kind here outlined.

At a symposium on specifications for physical properties of malleable iron castings held under the auspices of the American Society for Testing Materials, upwards of 20,000 test results[10] were contributed by seventeen different investigators.

Confining our attention to tensile strength it is of interest to note the graphical presentation of the ranges, Fig. 3, (the original test results not being available, we must make use of the ranges).

[10] Cf. Proceedings of the American Society for Testing Materials for 1931.

On one hand, let us assume that the inherent cost per casting is the same for each of the 17 sources. If so, why not raise the specification requirement from 50000 lb. per sq. in. (as specified in A.S.T.M. Designation A47-30) to some 54 or 55000 lb. per sq. in.? Why use source number 1 if material from number 5 is just as cheap? On the other hand, assume, as is more likely to be the case, that the inherent costs are not the same. Then for a given job there is a most economical form of the material, the specification for which is the economic standard. Enough has been said to indicate that the differences in the seventeen

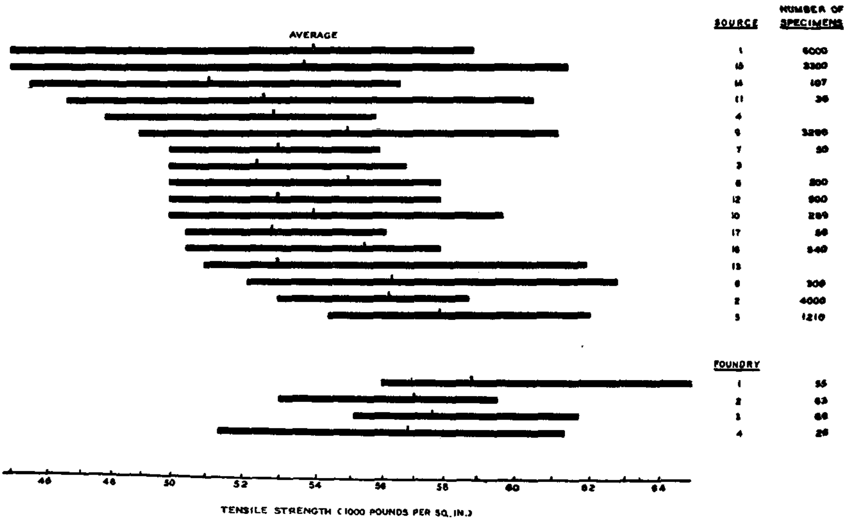

FIGURE 3. RANGES OF TENSILE STRENGTH FROM 22 SOURCES.

sets of data are likely greater than should be left to chance. *What is needed is a practical test for significant differences of this character—just what statistical theory provides.*

Only after discovering the assignable causes of the significant differences are we in a place to say what variability it is economical to accept in the standard, for only then do we know how much it will cost to take it out so that we can compare this cost with the increased value of the product resulting from the removal.

6. ECONOMETRICS AND INDUSTRIAL STANDARDIZATION[11]

In the preceding discussion an economic standard is conceived as a standard such that under the given conditions in respect to the development of science and the development of human wants, there is a balance between the economic value to the consumer of any possible

[11] This section was added September, 1932.

modification in the standard and the cost of such modification. This point of view is consistent with the vision of science in which there is not so much a struggle between men, or companies, or nations, for a limited store, where one's gain must be another's loss, as there is co-operation in an effort to raise the standards of living of all by making use of the results of progress in pure science.

It is obvious that in this effort there is need for the development of a *quantitative economic theory* which will take into account not only demand and supply but also quality. By nature the engineer is one who likes to set a goal and then try to attain it. But to do this, it is necessary for him to have available quantitative equations of economic theory in the same sense that he has such equations in the natural sciences—hence the significance of scientific economics, as supported by the Econometric Society, to the industrial leaders charged with making use of physical laws and properties in satisfying human wants.

Bell Telephone Laboratories