

R480  
WMD

I. E. B. 7

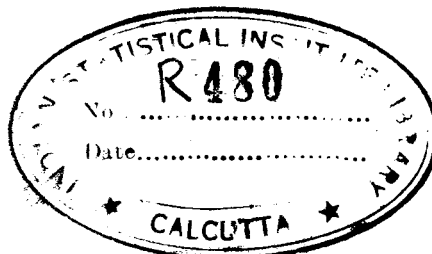
STATISTICAL METHOD FROM THE VIEWPOINT OF QUALITY CONTROL

W. A. SHEWHART'S COLLECTION

by

W. A. Shewhart

A report outlining the results of some recent studies in the development of an operationally verifiable statistical methodology and presenting a critical discussion of some of its potential contributions that are fundamental in: a) the attainment of economic control of quality, b) the establishment of tolerance limits, c) the presentation of data in the most useful form, and d) the specification of requirements as to accuracy and precision. This methodology is fundamentally different from that based upon the so-called modern statistical theory of inference in that: it starts with the assumption that a state of statistical control does not exist in general, instead of with the assumption that it does exist; it leads to prediction in terms of tolerance limits instead of fiducial limits; and it takes into account the significance of tests for both consistency and reproducibility instead of only the latter - three requirements imposed by the practical problem of quality control.



Bell Telephone Laboratories, Inc.  
Inspection Engineering Department

Issued for use of members of the Department

December 1937

I would almost say: "Show me a uniform product and I will show you a manufacturer who really understands his processes." There are two kinds of manufacturing psychology. One which cannot apply refinements and careful manufacturing control because of cut market prices. The other which must apply refinements and controls because prices are so cut. The wastage and improvisations which are the accompaniments of variability are far more expensive than the knowledge and controls which result in uniformity. It requires a genius to produce uniformity, but even a politician can be an expert in inconsistency.

C. C. PATERSON, Director  
Research Laboratories  
General Electric Co., Ltd.

TABLE OF CONTENTS

	<u>Page</u>
<u>Introduction</u>	
<u>Chapter I</u>	
Statistical Control . . . . .	1
<u>Chapter II</u>	
How Establish Limits of Variability . . . . .	24
<u>Chapter III</u>	
Presentation of the Results of Measurement of Physical Properties and Constants . . . . .	52
<u>Chapter IV</u>	
Specification of Accuracy and Precision . . . . .	83

## INTRODUCTION

"Interchangeable manufacture gives an output better in quality, cheaper in cost and more useful than would be possible without it. It constitutes one of the greatest contributions of the machine age ....."<sup>1</sup>

J. W. ROE, Professor of Mechanical Engineering  
New York University.

The above statement is made in a very interesting article which leaves off at the point where the present story begins, namely with the introduction of statistical techniques into control theory and practice. It is pointed out that with the introduction of these newer techniques, mass production of interchangeable parts takes a big step forward in achieving its goal of lower cost and greater usefulness of manufactured goods. Furthermore, such theory and technique is of necessity characteristically different in certain fundamental ways from that developed for use in fields of research. Hence there is what we may call statistical method from the viewpoint of quality control in mass production of interchangeable parts. The development of statistical methodology from this viewpoint is most likely in its infancy but progress has been made to such a point that one can at least set down some of the basic contributions that such methods make in the attainment of maximum advantages inherent in the process of mass production.

By interchangeable manufacture is usually meant the production of complete machines or mechanisms, the corresponding parts of which are so nearly alike that they will fit into any of the given mechanisms. The term fit as here used is likely, however, to suggest mechanical or electrical fits as contrasted, for example, with the substitutability from the viewpoint of quality of one part for any other part supposed to be the same. In this broader sense, one piece of a raw or fabricated material should be interchangeable in a quality sense with any other similar piece of the same material. Such a re-

-----  
1. Mechanical Engineering, October, 1937, pp. 755-758.

quirement of interchangeability applies not only to engineering materials but also to foods, drugs and the like.

Schematically we may represent any set of objects such as similar pieces or quantities of a raw or fabricated material by the symbols

$O_1, O_2, \dots O_1, \dots O_N, O_{N+1}, \dots O_{N+1} \dots$

What is wanted from the viewpoint of interchangeability may then be thought of as the requirement that one of these parts is for the purposes in question just as satisfactory as any other. If it were feasible to make the physical objects symbolized by the O's identical one with another in respect to all quality characteristics, then the O's would obviously be interchangeable in use.

Since it was not possible to attain this degree of conformance, the concept of a go-no-go tolerance was introduced sometime around 1870. For each quality characteristic, the requirement was made that this characteristic for each of the objects, represented by the O's, should lie within specified tolerance limits,

$$X = L_1 \text{ to } X = L_2.$$

Thus far, statistical concepts played no part. In fact it was not until years later that the need for statistical methodology became apparent. This came in trying to attain the most economic plan of production within tolerances. Our story begins at this point.

The first chapter sketches briefly the rôle of statistics in the three steps of controlling quality - specification, production and inspection - corresponding respectively to legislative, executive and judicial acts. We see why, for economic and quality assurance reasons it is necessary to introduce the concepts of aimed-at value and action or control limits and how these may be provided through the application of statistical theory. In this way we come to see the practical importance of the concept of a state of statistical control or homogeneity as a conceptual limit to which one may hope to go in making efficient use of materials and piece parts.

It turns out that the three steps in the control process are correlated in such a way that they cannot be taken independently and that the ultimate goal can only be reached through mass production. This comes about

because it appears that the physical state of statistical control is only approached gradually. Moreover, this fact is of marked significance from the viewpoint of statistical methodology in that it indicates the rather extensive investigations that apparently must be made before one can place much reliance on inferences from samples based upon the assumption that the sample has arisen from a state of statistical control. This chapter also considers briefly three important concepts of statistical control and shows the rôle that these play in specifying, producing and judging quality.

In the first chapter we start with the assumption that tolerance limits have already been established. In Chapter II we take up the problem of establishing such limits for both controlled and non-controlled statistical states. In many cases all one has to rely upon is the tabulated values of physical properties in tables of physical constants which incidentally are often given in the form

$$X \pm \Delta X$$

for any given quality  $X$ . Obviously in establishing tolerance limits it is often of great advantage to make the tolerance range a minimum. The second chapter, therefore, considers the problem of establishing limits of variability. In the first place, we find that theoretically this can only be done with the maximum degree of assurance after a state of statistical control has been attained. What is perhaps more important is that even under statistically controlled conditions the establishment of a tolerance range is a fundamentally different problem from establishing a range of the type usually discussed in statistical texts and books on the theory of errors. One can at least under idealized conditions set up a range of the form  $X \pm \Delta X$  that has an operationally verifiable meaning no matter how small the sample size, so long as it is not less than let us say three, and a range for a small sample is just as valid as a range for a larger sample. The interesting point is, however, that the meaning of the two ranges is fundamentally different. Satisfactory tolerance ranges can be set only upon the basis of comparatively large samples. This fundamental difference between tolerance ranges and the fiducial ranges of statistical theory has to the best of my knowledge not been previously discussed in the literature. Incidentally it is shown that some scientists make huge errors by confusing the two ranges. If as is found to be

the customary case, the variability does not satisfy the criteria of control, then the establishment of a tolerance range becomes a much more difficult problem and the best approximation involves not only the application of statistical techniques but also tests for logical consistency with other pertinent data.

We are more or less naturally led to the problem considered in the third chapter - the presentation of measurements of physical properties and constants. Thus in establishing tolerances we must go from a set of observed values to an estimated tolerance. Likewise in judging quality - the third step in the control process - we have this same general problem of tabulating great masses of data in a form of quality report that will be most useful. This is an exceedingly important problem - in fact it is also involved in any attempt to summarize the results of research as, for example, in tables of physical and chemical constants. This problem is usually discussed in the literature upon the basis of assumptions that do not hold in practice and hence is not directly applicable. The discussion of this subject must, therefore, take us beyond customary theory and practice if it is to be applicable. Some new ground is broken in this chapter that has a field of application beyond that of quality control. Here again we are guided not only by statistical tests but also by tests of consistency.

In order to attain any of the advantages of interchangeability it is necessary to satisfy three kinds of criteria - one for statistical control, one for precision, and one for accuracy. The fourth chapter tackles the problem of specifying requirements of accuracy and precision. In the first place, it is necessary to consider the confusion of these two terms in the literature and to show why it is very important in the theory and practice of quality control to differentiate between them. Moreover, it is shown that before we can place much significance on any quantitative measure of precision we must first have adequate evidence that a physical state of statistical control exists. In this chapter we are introduced to a criterion of operational verifiability and a differentiation between theoretical and practical verifiability, the first of which plays an important rôle in giving meaning to the "intent" of specifications and the second plays the role of criterion in the preparation of inspection specifications. Much of the discussion as to accuracy and precision has a broad bearing in many fields of science outside of quality control.

In brief, Chapter I starts where "exact" science left off with a concept of tolerance limits and shows how and why it is necessary to apply statistical techniques in putting two action limits and an aimed-at value between these tolerance limits. Chapter II starts a step back of this, namely, with the consideration of ways and means of establishing tolerance limits. Chapter III considers the problem of summarizing data from the viewpoint of testing hypotheses, so important in setting tolerances and in providing an adequate running quality report. Chapter IV closes with a consideration of the all important requirements of mass production of interchangeable parts - accuracy and precision and the specification of these in operationally verifiable terms. A consideration of these four problems serves to reveal the rapidly expanding field of application of statistical theory and technique in the control of quality in an economical way and in a way to provide maximum quality assurance. Incidentally some of these applications reveal hitherto undescribed but important characteristics of applied statistical methodology.



# Statistical Methods from the Viewpoint of Quality Control

## INTRODUCTION

1. Appreciate the honor.
  2. Unique experience in many ways.  
Majority of you said ~~no statistics~~.  
previous experience with business methods.
  3. You have unique experience listening to one who has had as much or more criticism of statistical methods. → My comments informal
  4. Limitations of method.  
→ Cowboy in Texas Saloon - evasion.
  5. Importance of homogeneity (control)  
Dr. John Johnson.
- Historic stages in control. -

$S_0$  | define limits A and B  
Expected values  $\bar{x}$  and  $\sigma$

### STATE of Control.

Physical { a) Same essential conditions.  
b) C System of chance causes (constant)  
c) state that will give Random sequence

Mathematical

$$dy = f(x, \theta_1, \theta_2, \dots, \theta_i, \dots) dx$$

How safe from one state to other  
Bridge between Phys. and the  
 How far from  $\theta$  to  $\theta'$   
 $S_7$

When is Bridge safe?

Assumption - Bowl.

Assumption Criterion of Action  $S_8$

to make for drawing for bowl  $S_9$

Is it safe for resistances above  
 slide

$S_{10}$

Is it safe for velocity of light

$S_{11}$

III Operation of Control Criterion

A Concept of Random distribution of  $n$

$S_{12}$

B Random order

$x_1$	$x_2$	$\dots$	$x_i$	$\dots$	$x_n$	$x_{n+1}$	$\dots$	$x_{n+i}$
$c_1$	$c_2$		$c_i$		$c_n$	$c_{n+1}$		$c_{n+i}$

Resistances not controlled

$S_{13}$

C Same first - no order.

$S_{14}$

D How attain State.

Remove assignable causes.

Then get 25 to 250 points  
within control.

Can be done

S15

## Judgment of Control

Past Experience  
is evidence

Prediction within  
limits

E

degree of belief  $p_b$

## Three Concepts of prob.

$$P_1 = \int_{x_1}^{x_2} f(x) dx$$

$p$  is same  $\lim_{n \rightarrow \infty} p = p'$

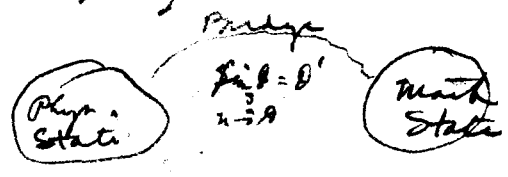
$p_b$  is same of judging.

## Three Steps operational production (operation) judgment

S16

Fig 4 on board  
Quantity  
 $Q_1, Q_2, \dots, Q_i, \dots, Q_{N+1}, Q_{N+2}, \dots, Q_{N+i}, \dots$   
 $X_1, X_2, \dots, X_i, \dots, X_N, X_{N+1}, \dots$   
 $\downarrow \quad \downarrow$   
 $C_1 \quad C_2$

$$dy = f(x_1, \theta_1, \theta_2, \dots) dx$$



## CHAPTER I

### STATISTICAL CONTROL




The possibility of improving the economy of steel to the consumer is therefore largely a matter of improving its uniformity of quality, of fitting steels better for each of the multifarious uses, rather than of any direct lessening of its cost of production.

1  
JOHN JOHNSTON, Director of Research  
United States Steel Corporation

There are three senses in which statistical control may play an important part in the control of quality of manufactured product. These are: (a) as a concept of a statistical state constituting a limit to which one may hope to go in improving the uniformity of quality, (b) as an operation or technique of attaining uniformity and (c) as a judgment. Here we shall be concerned with an exposition of the meaning of statistical control in these three senses and of the rôle that each sense plays in the theory and technique of economic control. But first we should consider briefly the history of the control of quality up to the time that engineers introduced the statistical control chart technique which is in itself an operation of control.

#### SOME IMPORTANT HISTORICAL STAGES IN CONTROL OF QUALITY

To give us a perspective from which to view recent developments, let us look at Fig. 1. That which to a large extent differentiates man from animals is his control of his surroundings and particularly his production and use of tools. Apparently the human race began the fashioning and use of stone implements about a million years ago as is evidenced by the crude stone implements shown to the left of Fig. 1 which were recently discovered just north of London.<sup>2</sup> Little progress in control was made, however, until about 10,000 years ago when man first began to fit parts together as evidenced by the holes in the instruments of that day shown in Fig. 1.

1,000,000 YEARS AGO	150,000 YEARS AGO	10,000 YEARS AGO	150 YEARS AGO
			INTRODUCTION OF INTERCHANGEABLE PARTS

Throughout this long period ap-

Fig. 1

1. "The Applications of Science to the Making and Finishing of Steel", Mechanical Engineering, February, 1935.
2. This discovery is reported in Man Rises to Parnassus by H. F. Osborne, Princeton University Press, 1928. The photograph of the stone implements of a million years ago has been reproduced by permission from this most interesting book. The implements of 150,000 to 10,000 years ago have been reproduced by permission from the fascinating story told in Early Steps in Human Progress by H. J. Peake, J. B. Lippincott and Company, 1933.

parently each man made his own tools, such as they were. As far back as 5000 years ago the Egyptians are supposed to have made and used interchangeable bows and arrows to a limited extent. It was not, however, until about 1787 or a hundred and fifty years ago that we had the first real introduction of the concept of interchangeable parts. Only yesterday, as it were, did man first begin to study the technique of mass production!

From the viewpoint of ideology it is significant that this first step was taken under the sway of the concept of an exact science. Accordingly an attempt was made to produce piece parts to exact dimensions. How strange such a procedure appears to us today, accustomed as we are to the concept of tolerances. But as shown in Fig. 2, it was not until about 1840 that the concept of a "go" tolerance was introduced and not until about 1870 that we find the "go no-go" tolerance.<sup>1</sup>

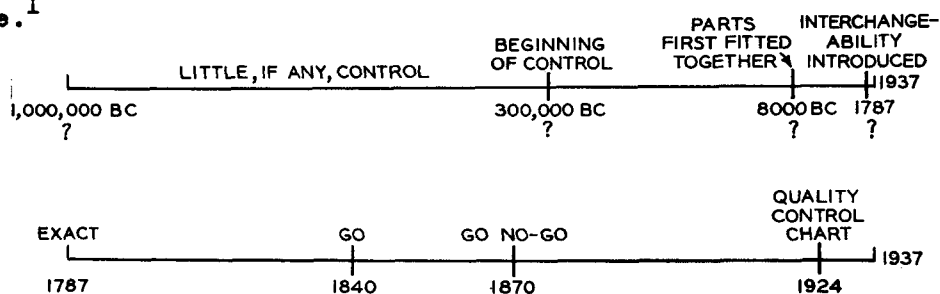


Fig. 2

Why these three steps: exact, go, go no-go? The answer is quite simple. Manufacturers soon found that they could not make things exactly alike in respect to a given quality, it was not necessary that they be exactly alike, and it was too costly to try to make them alike. Hence by about 1840 they had eased away from the requirement of exactness to the go tolerance. Still too much time was wasted unnecessarily in trying to stay reasonably close to the tolerance. Then came the idea of specifying the go no-go tolerance or the range within which the quality characteristic might vary and still be satisfactory. This was a big forward step because it gave the production man more freedom and brought a still greater reduction in cost. All he had to do was to stay within the tolerance range - he didn't have to waste time trying to be unnecessarily exact.

1. It will be noted that the first six dates shown in Fig. 2 are given with a question mark - authorities are not in unanimous agreement as to the exact dates. I think, however, that the dates here shown will be admitted by all to be approximately correct.

Though this step was of great importance something else remained to be done. The way the limits are necessarily set is such that every now and then pieces of product are produced with a quality characteristic falling outside the specified range - in other words - defective. To junk or modify such pieces adds to the cost of production. But to find the unknown or chance causes of defectives and try to remove them also costs money. Hence after the introduction of the go no-go tolerance there remained the problem of trying to reduce the fraction  $p$  of defectives to a point where the rate of increase in cost of control equals the rate of increase in the savings brought about through the decrease in the number of rejects.

For example, in the production of the apparatus going into the telephone plant, raw materials are gathered literally from the four corners of the earth. More than 110,000 different kinds of pieceparts are produced. At the various stages of production inspections are instituted to catch defective parts before they reach the place of final assembly and are thrown out. Here one finds the problem of determining the economic minima for the sizes of the piles of defectives thus thrown out.

This problem of minimizing the per cent defective, however, was not the only one that remained to be solved. Tests for many quality characteristics - strength, chemical composition, blowing time of fuse, and so on - are destructive. Hence every piece of product cannot be tested for such a characteristic to see if it falls within the specified tolerances. Engineers must appeal to the use of a sample. But how large a sample should be taken in a given case in order to give adequate quality assurance?

The attempt to solve these two problems, giving rise to the introduction of the quality control chart technique in 1924, may therefore be taken as the starting point of the contributions of statistical technique to the control of quality of manufactured product in the sense here considered.

#### Why after 1900?

Why, you may ask, was it something like one hundred and fifty years from the start of mass production of interchangeable parts to the time of the more or less intensive study of the application of statistical methods in this field? There are at least two important reasons.

First, there was the rapid growth in standardization. Fig. 3 shows the rate of growth in the number of industrial standardization organizations both here and abroad. The first one was organized in Great Britain in 1901. Then beginning in 1917 we get a rapid spread of the realization of the importance of national and even international standards. Fundamentally the output

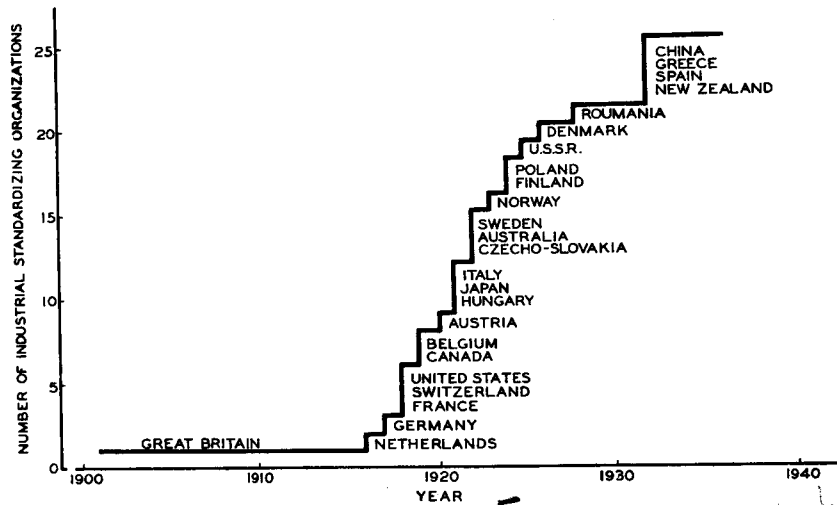


Fig. 3 53 #15426

of such standardization organizations is specifications of the aimed-at quality and in certain instances of methods of measuring this quality. But when one comes to write such a specification, he runs into the two kinds of problems - minimizing the number of rejections and minimizing the cost of inspection to give an adequate degree of quality assurance - discussed in the previous section. Hence the growth in standardization spread the realization of the importance of such problems in industry.

Second, there was a more or less radical change in ideology. We passed from the concept of the exactness of science in 1787, when interchangeability was introduced, to probability and statistical concepts which came into their own in almost every field of science after 1900. Whereas the concept of mass production of 1787 was born of an exact science, the concept underlying the quality control chart technique of 1924 was born of a probable science.

We may for simplicity think of the manufacturer's trying to produce a piece of product with a quality characteristic falling within a given tolerance range as being analogous to shooting at a mark. Now, if one of us were



shooting at a mark and failed to hit the bull's-eye, and some one asked us why, we would likely give as our alibi, CHANCE. Had some one asked the same question of one of our earliest known ancestors, he might have attributed his lack of success to the dictates of fate or to the will of the gods. I am inclined to think that in many ways one of these alibis is just about as good as another. Perhaps we are not much wiser in blaming our failures on chance than our ancestors were in blaming theirs on fate or the gods. But since 1900 the engineer has proved his unwillingness to attribute all such failures to chance. This represents a remarkable change in ideology which characterizes the developments in the application of statistics in the control of quality.

Starting with the introduction of the go no-go tolerances of 1870, it became the more or less generally accepted practice to specify for any given quality characteristic X that this quality should lie within specified limits  $L_1$  and  $L_2$ , represented schematically in Fig. 4.

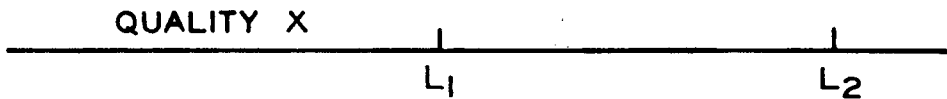


Fig. 4

Such a specification is of the nature of an end requirement on the specified quality characteristic X of a finished piece of product. It provides, as it were, a basis on which the quality of a given product may be gauged to determine whether or not it meets the specification. From this viewpoint, the process of specification is very simple indeed. Knowing the limits  $L_1$  and  $L_2$  within which it is desirable that a given quality characteristic X shall lie, all we need to do is to put these limits in writing as a requirement on the quality of a finished product. With such a specification at hand, it is presumed to be possible through measurement to classify a piece of product as conforming or non-conforming to specification.

As we have seen, however, two difficulties arise with this form of specification. Suppose that the quality in question, the blowing time of a fuse for example, is one that can be determined only by destructive tests. How can one give assurance that the quality of a given piece of product will meet the specification without first destroying it? Or again, even where the quality characteristic may be measured, there is always a certain expected

*So called process control*

*Reduce sample size  
Reduce rejection  
minimize variability*

fraction p falling outside the tolerance limits. How can we go about attaining an economic minimum to this fraction non-conforming? A little reflection shows that the simple specification of a go no-go tolerance is not satisfactory in such instances from the viewpoint of: 1) Economy, and 2) Quality Assurance.

*Tolerance*

The story of how statistical techniques can be used successfully to attain economies and to provide maximum quality assurance has been told elsewhere.<sup>1</sup> This we shall consider as water over the dam and start here with a consideration of control from the viewpoints of specification, production, and inspection of quality. This brings us at once face to face with the three senses in which the phrase "statistical control" may be used. To illustrate, suppose we fix our attention on some kind of material, piece part or physical object which we wish to produce in large quantities. Let us symbolize the pieces of this product by the letters

$$O_1, O_2, \dots O_1, \dots O_N, O_{N+1}, \dots O_{N+1}, \dots \quad (1)$$

Given a process of production, it presumably may be employed to turn out an indefinite number of pieces. Now let us consider the requirement:

- A. The quality of the O's shall be statistically controlled in respect to the quality characteristic X.

As an example, the O's might be condensers, and the X, a capacity; they might be pieces of steel and the X might be carbon content; or they might be any other kind of object and associated quality characteristic. Alongside this requirement let us consider also the meaning of the statement:

- B. The quality of the O's is controlled in respect to the quality characteristic X.

On the face of it, the only difference between A and B is that in one the verb "shall be" is replaced by "is". The requirement A, however, is of the nature of a characterization of a state of control whereas B is a judgment which may or may not be true. The process or operation of producing the O's is the operation of control and may involve the use of statistical techniques such as that of the control chart. Let us, therefore, examine each of these concepts of control in turn.

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

## THE PROBLEM OF QUALITY CONTROL

When one attempts to turn out pieces of product the quality characteristics of which will meet specified tolerance limits, he runs into two difficulties. First, as noted above, it is general practice to set tolerance limits for a particular characteristic in such a way that not all pieces of the product meet the tolerance requirements. Hence there is usually a fraction  $p$  of any lot of product turned out which does not conform to the tolerance requirements and which must therefore be given special attention. The added labor that this entails, as well as the fact that some or all of this non-conforming material may have to be junked, adds to the cost of production and hence there arises the problem of trying to reduce the fraction  $p$ , non-conforming to a minimum. In the second place, many of the specified quality characteristics cannot be inspected except by the use of destructive tests, - the blowing time of a fuse or the chemical content of some material, for example. Hence, without inspecting all of the product turned out, we must determine ways and means of giving maximum assurance that the quality of product, if and when tested, will be found to meet the specified tolerance requirements. These are two of the practical problems that originally led to a consideration of the application of statistical theory in quality control. In this section we shall try to formulate the problem of control from the viewpoint of these two problems.

In the first place, how is one to determine how far it is economically feasible to go in reducing the fraction non-conforming? To modify it for a given kind of product would evidently involve some modification of the manufacturing process. It is generally assumed that there is a limit to which one may hope to go in reducing the variability in the quality of a product, representing, as it were, the maximum degree to which one could expect to go in controlling quality. We might think of this as a state of maximum control of the physical cause system involved in the production process. It would obviously be a waste of money to try to reduce variability beyond this point; but even before it is reached, the economics of the production process may make it wise to consider that we have gone as far as it is economically feasible to go in reducing variations in a quality characteristic. However, the state of maximum control represents a fundamental limiting state of control. If in any case the process has not reached this state, there is something that we can do to modify the fraction defective without changing the whole process. But how are we to know whether or not the process has reached this state? This question obviously leads us to a consideration of a fundamental problem, namely, that of characterizing in an operationally definite way what we have termed the state of maximum control.

Next let us consider the problem of rendering maximum quality assurance under those conditions where, for economic or other reasons, it is necessary to rely upon the results obtained

from a sample. In such instances what we are interested in doing is to make a valid probable inference as to whether or not a piece or pieces of product not yet tested will, if and when tested, meet the tolerance requirements in respect to specified quality characteristics. Note that we introduce here the term "valid". It goes without saying that any one can make a prediction as to the happening of some event in the future, such as the prediction that not more than a certain percentage of a given lot of product will be found non-conforming. But all predictions of this character will not be of equal validity. It is obvious that there are some conditions which we may set up wherein it seems to be possible to make valid probable inferences. In the throwing of a coin and certain problems of drawing chips from a bowl, for example, it appears that we may use probability theory as a basis for making valid predictions. On the other hand, there are many problems in economics and the social sciences, and even in the physical and chemical sciences, where it does not seem feasible to make predictions upon the basis of probability theory with anything like the validity which one may expect in the case of drawings from a bowl.

Now if we are to render maximum quality assurance in a given case, we should <sup>try</sup> attempt to reduce to a minimum errors involved in prediction. In other words, we must attain a state of control of the cause system such that we may be justified in applying the statistical theory of probability as a basis for prediction. But if we are to do this, we must

find in some way or other operationally verifiable means of characterizing such a state of statistical control.

Let us assume that we have the specification for some particular kind of apparatus, fabricated material, piece-part, or physical object, which we wish to produce in large quantities. Suppose that we have before us for consideration a given process of production capable of turning out an unlimited number of pieces of the given kind of product. We may symbolize these by

$$O_1, O_2, \dots O_N, O_{N+1}, \dots O_{N+1}, \dots \quad (1)$$

It should be noted that the description given above of a state of maximum control and also of a state of statistical control refers to the state or condition of the cause system in the production process. In fact, any attempt to control leads one to think of this cause system as the means by which our control is effected. The object of attaining control, however, has been stated in terms of the product turned out by the process of production.

### STATISTICAL STATE OF CONTROL

#### Physical State of Statistical Control

Let us start with the concept of state of maximum control mentioned in the previous section. How is one to know when such a state has been attained? This concept seems to be related quite closely to another which is described by the phrase "the same essential conditions" in the literature

of exact science. Now the phrase, same essential conditions, seems to be applied wherever the experimentalist decides more or less intuitively that he has gone as far as he can in finding and removing causes of variability in his results. Such a criterion, however, cannot serve to characterize in an objective way the state of maximum control. In fact, the test is subjective and depends upon the experimentalist, as is evidenced by the fact that usually not all authorities will agree that all conditions have been maintained essentially the same.

There is a certain type of phenomena, however, where most authorities do agree that fluctuations in the observed phenomena must be left to chance or unknown causes. I have in mind the drawings of numbers from a bowl. Thus, if we have, let us say  $M$ , physically similar chips, and if we write some number on each of the  $M$  chips, place the chips in a bowl, and draw successive samples of  $n$  chips one at a time with replacement and thorough mixing, I think there will be almost unanimous agreement that the complexion of such samples is beyond human control. Such a series of drawings perhaps approach as closely as we can go today in characterizing in an operational way an example of what we have termed a physical state of maximum control.

This fact early suggested that experiments be performed to determine whether or not sequences of results of repeating such operations as making measurements and producing pieces of product under the same essential conditions satisfied certain criteria that drawings from a bowl satisfy. In almost

every instance among a very large number of trials, negative results were obtained. In such cases it was usually found possible to find and remove one or more assignable causes of variation until the resultant variation seemed to satisfy the criteria satisfied by drawings from a bowl. For such reasons, therefore, fluctuations in samples drawn from a bowl of chips seem to give a means of characterizing the state of maximum control in terms of the observable results produced by such a state.

Next let us consider the problem of characterizing the conditions under which we may expect to make valid predictions by means of probability theory. For some time at least, particularly in the theory of errors, there has been a tendency to assume that the application of probability theory is justified when the experimentalist has reached the state which he describes by the phrase "the same essential conditions". Likewise, much of small sample theory in modern statistics seems to imply the assumption that when the experimentalist has obtained his data under presumably the same essential conditions, he may without hesitancy make predictions as though the cause system had been reduced to a state of statistical control.

It should be noted that in general no requirements are explicitly stated in such applications of error theory and small sample theory as to how many observations are to be made prior to concluding that the conditions are essentially the same, nor are there in general any criteria imposed upon the



data - particularly if they are few in number - except that they represent observations that the experimentalist judges to have been made under the same essential conditions. This is pretty much like saying that when the experimentalist has gone as far as he thinks he can go, then he can expect to make valid predictions by probability theory.

In quality control work it has been found that we are not, in general, justified in assuming that predictions made under such conditions will prove valid. In fact, they are likely to be considerably in error. In the following chapters this same conclusion seems justified, even in the case of some of the most refined physical measurements. Hence it is that quality control engineers have been forced to seek further for conditions characterizing the physical state of control than any subjective criterion that the data have been obtained under the same essential conditions, in order to know when to expect applications of probability theory to lead to valid conclusions.

The drawings from a bowl, however, also serve as an example of an experiment carried on under a physical state of statistical control, in the sense that such drawings produce results which satisfy the probability laws of statistics as closely as any experimental results known today. These facts suggest that unless fluctuations in an observable phenomenon satisfy criteria satisfied by drawings from a bowl, one is not justified in applying probability theory.

What is more important, however, is the fact that the observed data must not only satisfy certain criteria but

that there must be not less than a certain quantity of data available.

February 3, 1938

Mathematical State of Statistical Control

By this phrase I shall refer to the abstract mathematical theory used in describing the observable phenomena characterizing the idealized physical state of statistical control. For example, let us consider a production process which is capable of turning out an indefinitely large number of objects which we may symbolize by

$$O_1, O_2, \dots O_1, \dots O_N, O_{N+1}, \dots O_{N+1}, \dots O_{2N}, O_{2N+1}, \dots O_{2N+1}, \dots O_{kN} \dots$$

Let us consider successive lots of size  $N$  and let us assume that the number of pieces non-conforming in these lots are respectively

$$p_1N, p_2N, \dots p_iN, \dots p_kN \dots \quad (2)$$

One of the simplest and yet most fundamental problems is to characterize the distribution of such a sequence of values found non-conforming in a way that we shall choose to say represents the idealized physical state of statistical control. This is done, of course, by saying that in such a state the probability of finding  $0, 1, 2, 3, \dots N$ , conforming in a lot of  $N$  is given by the terms of the point binomial

$$N(p'+q')^N, \quad (3)$$

where  $p'$  is some unknown constant and where  $p' + q' = 1$ . The constant  $p'$ , of course, is termed the mathematical probability of a piece of product conforming in such a case

and  $q'$  is the probability of a piece of product not conforming.

We may, however, approach this problem in a little different way. To illustrate, let us confine our attention to a single quality characteristic. For example, let the sequence

$$X_1, X_2, \dots, X_i, \dots, X_N, X_{N+1}, \dots, X_{N+i}, \dots \quad (4)$$

represent the observed values of a quality characteristic of the infinite sequence of pieces that a given process may turn out. We may then ask ourselves for a characterization of such a sequence of numbers corresponding to an idealized physical state of statistical control. One way, of course, of approaching this problem is to assume that there is for such a state in a given case a characteristic but unknown function  $f(X)$  such that the mathematical probability  $dp$  that a piece of product will have a quality characteristic  $X$  lying within the interval  $X \pm 1/2dX$  is given to a high order of approximation by the expression

$$dp = f(X)dX. \quad (5)$$

In practice, however, all that one ever has is a sequence of observed values of the form of either (2) or (4), whereas the theoretical distribution functions (3) and (5) characterize corresponding infinite sequences.

Now, of course, if one knew that he was dealing with a situation that could be characterized by the distribution theory associated with a characteristic function such as either the point binomial or a continuous frequency distribution, he could then use distribution theory to calculate in

a mathematically rigorous way the nature of sampling fluctuations that might be expected in samples of any given size and for an unlimited number of different kinds of functions of the numerical values contained in the samples. We are perhaps justified in considering that such distribution theory provides us with an indefinitely expansible means of characterizing so-called sampling fluctuations, derivable in terms of formal rules of mathematics. The nature of the distribution function, however, is known to depend, in general, upon the functional form  $f$  of the characteristic distribution function. Furthermore, the form of the function is unknown as well as the parameters. What is far more important, however, from the viewpoint of our present discussion is the fact that we do not know, in any given case, whether or not probability theory is applicable in the case at hand. This leads us naturally to our next point of considering the customary method of explaining how one may hope to relate, as it were, a given physical state and the associated ideal characteristic distribution function such as either (3) or (5), for example.

Statistical Limit as a Bridge between the Physical and Mathematical States of Statistical Control

Let  $\theta_i$  be a parameter in the equation of the characteristic distribution function. It is possible to find some statistic  $\theta'_i$  of, let us say the first  $n$  values, of an observed sequence of measurements of quality such that the statistical limit of  $\theta'_i$  approaches  $\theta_i$  as  $n$  is made to approach infinity, or symbolically

$$\lim_{n \rightarrow \infty} \theta'_i = \theta_i$$

So far as I am aware, however, the concept of statistical limit cannot be used mathematically, at least in the sense of the ordinary concept of limit. What then is the significance of such a limit concept? Given any observed sequence representing the measurements of quality on a series of objects produced by a given process, it certainly is not possible to show that this sequence approaches in any accepted mathematical sense some limiting value. It may be helpful to consider an example of such a statistical approach to a limit.

STATISTICAL STATE OF CONTROL

Let us approach our discussion of the concept of state of control from the practical angle. We have already noted how engineers have gone from the concept of exact to the use of go, go no-go, and finally go no-go plus control chart for economic and quality assurance reasons. Fundamentally what is wanted is:

- a. A rational way of predicting that is subject to minimum errors, and
- b. Minimum variability in quality at a given cost of production.

We might then think of the ideal of uniformity of quality being characterized as that physical state satisfying requirements a and b. But how are we to know when such a state has been attained?

One might answer, of course, that this state is attained only by doing the best we can to control conditions of production so that they remain "essentially the same". All of us know how often this phrase enters into discussions of measurements in the so-called exact sciences. In fact, the phrase seems to be used to describe an idealized state which is tacitly assumed to satisfy requirements of the type a) and b) stated above.

Now in order to satisfy the condition of predictability, it is necessary to introduce some postulate involving probability. Hence it is but natural to think of a chance cause system controlling variation in quality in such a way that the probability  $dp$  of producing a piece of product with the quality  $X$  lying within the range  $X \pm 1/2 dx$  is to a high order of approximation given by an expression of the form

$$dp = f(x) dx. \quad (2)$$

where  $f$  is some mathematical function. Such a system of causes might be thought of as defining a statistical state of control which would satisfy requirement (a).

Error theory and much "small sample theory" in modern statistics seem to make applications which tacitly involve the assumption that when a scientist has obtained his data under presumably the same essential conditions he may without hesitation make predictions as though the conditions for equa-

tion (2) to hold were satisfied. There are, in general, no requirements as to how many observations are to be made prior to concluding that the conditions are essentially the same, nor are there any criteria imposed upon the data (if they are few in number) that are available except that they satisfy the condition that in so far as the experimenter knows they were taken under the same essential conditions.<sup>1</sup>

Thus far the statistician as a statistician has not contributed very much to the picture. True enough, equation (2) might be thought of as belonging to the field of statistics, but if one looks critically at the context in which it appears, he sees that it is used in a description of a physical chance cause system, the very nature of which is assumed to be unknown. In fact, so far as this system is thought to be synonymous with conditions which are essentially the same, the causes are supposedly unknowable.

The condition would be radically different if the statistician could write down a sequence of numbers which would characterize once and for all what a statistically controlled state of causes might be expected to give. Let us assume for a moment that this were possible and that

$$s_1, s_2, \dots s_i, \dots s_n, s_{n+1}, \dots s_{n+i} \dots \quad (3)$$

represented such an infinite sequence with which we might compare any observed sequence,

$$X_1, X_2, \dots X_i, \dots X_n, X_{n+1}, \dots X_{n+i}, \dots \quad (4)$$

The sequence (3) might then be thought of as a fairly definite quantitative description of a physical statistical state in terms of the sequence of numbers which that state would produce. This certainly would be a problem for the statistician.

With the same assumptions in mind, let us consider the problem of comparing an observed sequence with the standard sequence representing a statistically controlled condition. Theoretically we might conceive of ways of

-----  
1. Of course, if they are few in number, little else could be done, but the trouble is that there should be some requirement upon the number to be taken, as we shall see in the next chapter.



comparing the whole of an observed sequence with the standard. What is wanted from a practical viewpoint is, however, a criterion or set of criteria that the first  $n$  members of an observed sequence must satisfy in order to be reasonably sure that the remainder of the sequence will have the characteristics of the standard sequence (3). This problem is fundamentally different from that of constructing the standard sequence (3) which is a problem for the mathematical statistician. However, the justification of any proposed solution of the problem of setting up criteria that will work reasonably well on the basis of an examination of the first  $n$  terms of an experimental sequence must in the last analysis rest upon empirical evidence. The field of quality control in mass production offers undoubtedly the best "proving ground" known today for testing out the validity of the chosen techniques in that the predictions of today are almost sure to be tested tomorrow, and those of tomorrow, the next day and so on indefinitely. Whether or not there have existed or may exist physical states of control that give a sequence resembling the assumed standard sequence can only be determined experimentally.

Now let us return to the statistician's problem of writing down the standard sequence (3) for comparison. Even though we assume a perfectly definite functional relationship in equation (2) there is no unique way of writing down a comparison sequence. There are, instead, an indefinitely large number of ways in which such a sequence may be characterized in terms of an indefinitely large number of different statistics and methods of breaking up the sequence into subsamples. In other words, we have here all of the results of what is generally termed the mathematical theory of distribution to which contributions are being added daily. Personally, I like to look upon the theory of distribution as providing an indefinitely large reservoir of criteria by which one might characterize a sequence which in turn we might think of as characterizing the observables of a physical state of statistical control. It should perhaps be added that the nature of the characterization provided by distribution theory depends upon the functional form  $f$  in equation (2). In this connection, it is fortunately true that the characteristics in terms of certain statistics, such, for example, as the arithmetic mean and the fraction within specified limits, is practically independent of the functional form  $f$ , at least over a

very wide range.

Before passing on to a consideration of the operation of approaching the assumed ideal state of statistical control, let us consider another very important characteristic of the concept of the distribution or sequence given by such a state. We say that under such conditions we may find certain statistics of a sample of  $n$  which approach in the sense of a statistical limit,  $\text{Lim}_s$ , certain constant values as the sample size  $n$  is increased indefinitely. Thus for a given statistic  $\theta_1$  we say

$$\text{Lim}_s \theta_1 = \theta'_1. \quad (5)$$

$n \rightarrow \infty$

Even in the case of the existence of a statistically controlled state, we need to find the parameters in the functional relationship, and so far as we are here concerned the practical significance of the concept of statistical limit seems to be that it implies that the only road to improvement in an estimate of a parameter through the use of even the so-called most efficient statistic is to increase the sample size  $n$  - that is, by the method of repetition. In the face of this, however, statisticians as a rule neglect to tabulate the sample size  $n$ , the importance of which we shall see in the later chapters.

So far as I am aware, the concept of statistical limit cannot be used mathematically and in fact constitutes a non-mathematical characteristic of the state of control. In this sense it constitutes pretty much, as it were, the postulate of an operational rule of inference depending upon sample size. That is to say, if it were feasible to find a standard sequence (3), it would still be necessary to add the postulate of statistical limit to characterize what is customarily meant by a state of statistical control.

Physically perhaps the nearest approach one can get to the nature of a statistical limit is with drawings with replacement from an experimental universe written on a series of "physically similar" chips. Fig. 5 is one such observed approach which may serve to illustrate some of the points which can well be emphasized in trying to get at an understanding of the concept of statistical limit as here presented. The distribution in the bowl was approximately normal and symmetrical about zero as an arithmetic mean. The ordinate

of each point is the observed average for the sample of size corresponding to the abscissa of that point. It is of interest to note how the observed average swings back and forth about zero which is sometimes spoken of as the theoretical limit.

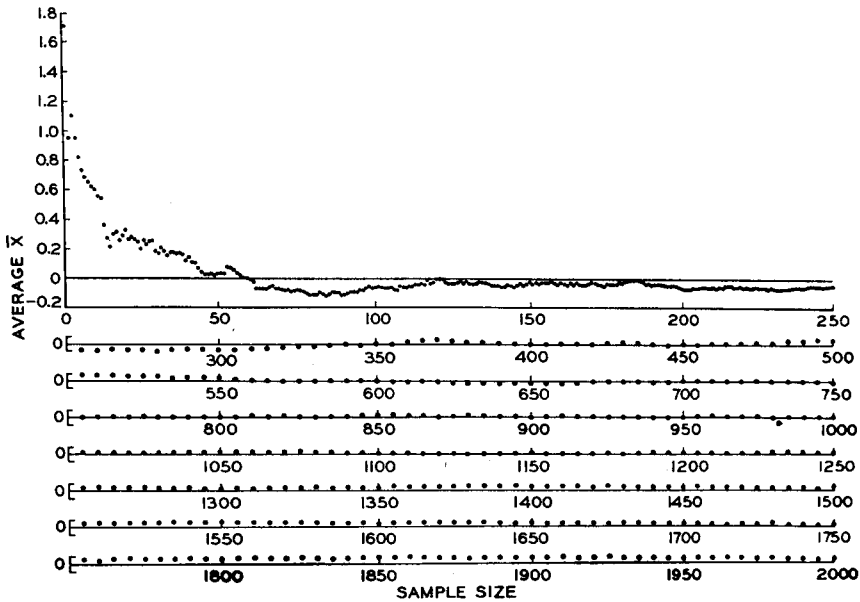


Fig. 5

Strictly speaking, zero would be the theoretical limit only if the chips were physically similar, a fact which we can never know with certainty. Do we know that this average approaches some value  $\bar{X}$  in the sense of a statistical limit in this particular case? No matter how many observations we take, I do not know how we could answer this question with certainty.

One might ask if this approach satisfies that symbolized formally by (5). I know of no way of checking this statement in an operationally definite way any more than I know of any way of checking once and for all a sequence in an operationally definite way to see if it represents a statistical state. If we assume that the dotted curve in Fig. 5 approaches a limit, the most practical significance of this conclusion is that we are tacitly adopting as a rule of operation that an average of  $n$  observations is to be taken in preference to  $n-1$ , let us say.

Of course, one might always wish to reserve judgment in a given case until he had compared the sequence of observed values with a set of criteria chosen arbitrarily to test whether or not the numbers are to be accepted as having arisen from a statistical state. But as noted above, there is no a priori and unique rule for choosing such a set of criteria. Hence in my own work, I prefer to say that drawings with replacement and thorough mixing from a bowl characterize a physical state of statistical control representing the limit

to which one may hope to go in attaining valid predictability and a state where the one making the drawings as prescribed cannot do anything to control the limits of observed variability - that is, it satisfies criteria a) and b) with which we started. It must, however, be kept in mind that logically there is no necessary or formal a priori connection between such a physical statistical state and the indefinitely expansible concept of statistical state in terms of mathematical distribution theory. There is, of course, abundant evidence of close similarity if we do not question too critically what we mean by close. What is still more important in our present discussion is that if this similarity did not exist in general and we were forced to choose between the formal mathematical description and the physical description, I do not see how we could get around looking for a new mathematical description instead of looking for a new physical description for the latter is what we apparently have to live with. It is the practical man's good fortune that mathematical distribution theory seems to check so closely what he gets in drawings from an experimental universe. As an indirect result, distribution theory must become the stock in trade of the control engineer.

The importance of considering in detail some of the more or less obvious points in this section will, I hope, become clearer as we proceed to discuss the other two senses - particularly that of a judgment - in which statistical control is to be considered.

#### STATISTICAL CONTROL AS AN OPERATION

In the beginning we noted the steps taken in going from the concept of an exact fit based upon the concept of an exact science to the concept of tolerances, Fig. 4.

At this point statistical theory stepped in with the concept of two limits A and B which we shall term action or control limits, and which lie, in general, within  $L_1$  and  $L_2$ . These limits are designed to be such that when the observed quality of a piece of product falls outside of them, even though the observation be still within the limits  $L_1$  and  $L_2$ , it is desirable to look at the manufacturing process in order to discover and remove, if possible, a cause of variation which need not be left to chance. In other words, whereas limits

Page 11, Second paragraph

What then is the force of the concept of statistical limit from the viewpoint of human action? So far as I see, it simply constitutes, as it were, a basis for a rule of action which in the absence of any better rule we accept as a method of acting under conditions where we assume that we have an ideal physical state of statistical control such as represented by a bowl. That is to say, if we accept the limiting process in such a case, it is pretty much equivalent to accepting, in general, a statistic  $\theta_1$  calculated from  $n+1$  observations in preference to the same statistic calculated from  $n$  observations.

Even though one were to accept such a method as being perhaps the best thing to do in the case of samples drawn from a bowl, he would still be left in the dark as to what to do in a given case where the very question which he is trying to answer is: Does the observed sequence arise from a physical state of statistical control? Are we justified, for example, in assuming that a statistic calculated from  $n+1$  observations is to be preferred over one calculated from  $n$  observations?

We might, of course, answer this question by saying that if the sequence appears to arise under the same essential conditions, then such a procedure is justified. But we have already called attention to the fact that experience shows that such sequences do not, in general, satisfy certain criteria that samples from a bowl satisfy. Hence it appears that the concept of a statistical limit as a bridge by which we may pass from the physical state of control to its mathematical

description does not serve to get us over the real difficulty of determining whether or not in a given case an observed sequence arises from a physical state of statistical control. In other words, we need to find some kind of an operation that is perfectly definite and verifiable that could be applied to an observable portion of an infinite sequence even when taken under presumably the same essential conditions that may be used to differentiate those cases where we may expect to make valid application of probability theory from those for which such application does not seem justified. This leads us, in other words, to a consideration of what we shall term the operation of attaining a state of statistical control.

$L_1$  and  $L_2$  provide a means of gauging product already made, action limits A and B provide a means of directing action toward the process in order that the quality of product not yet made may be less variable on the average.

Furthermore, the statistical theory of quality control introduces the concept of another point C lying somewhere between the action limits A and B which is the expected or in a certain sense the aimed-at value of quality in an economically controlled state. We should perhaps pause a moment to note the significance of the point C from the viewpoint of design or the use of material that has already been made. Let us take, for example, a very simple case of setting over-all tolerances. Suppose we start with the concept of the go no-go tolerance of 1870 and wish to fix the over-all tolerance for n pieceparts assembled in such a way that the resultant quality of the n parts is the arithmetic sum of the qualities of the parts. An extremely simple example would be the thickness of a pile of n washers. The older method of fixing such a tolerance is to take the sum of the tolerances on the pieceparts. This is generally many times too large from the viewpoint of economy. The efficient way of setting such tolerances is in terms of the concept of the expected value and the expected standard deviation about this value. In other words, the concept of expected value is of fundamental importance in all design work in which an attempt is made to fix over-all tolerances in terms of those of pieceparts.

Thus we see how, starting with the simple concept of a go no-go tolerance in a specification as illustrated in Fig. 4, it is necessary in many cases<sup>1</sup> for economy and quality assurance reasons to introduce certain action limits A and B and also a certain expected value C to be used in design formulae. The situation corresponding to the simplest case is shown schematically in Fig. 6. Statistical theory alone is responsible for the introduction

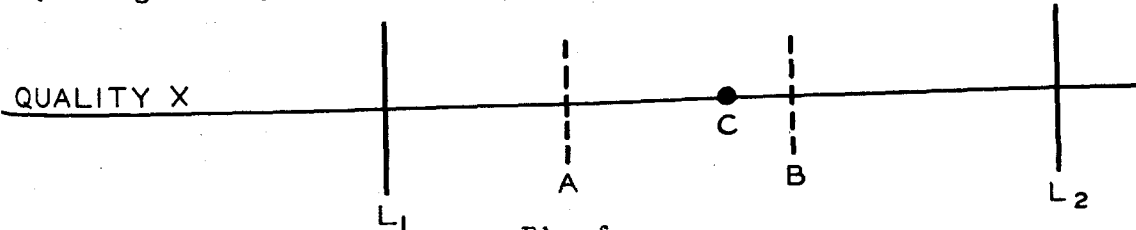


Fig. 6

<sup>1</sup>It should be noted, of course, that if there is no economic or quality assurance reason for going beyond the concept of the go-no-go tolerance, statistical theory has nothing to add. Likewise, it should be noted that although the action limits A and B may lie within the tolerance limits  $L_1$  and  $L_2$  product already produced and found within the limits  $L_1$  and  $L_2$  is still considered as conforming although outside A and B. In other words, the action limits A and B do not apply as a gauge for product already made.

of the concept of action limits A and B and the expected value C.

In the sense that the use of such statistical techniques introduces a modification in the operation of control, they constitute an "operation of statistical control". Their use serves as an operation directed toward attaining a state of statistical control in the sense of the previous section. They are means to the end of obtaining a state where mathematical distribution theory may be applied with the same assurance of validity as it can be applied to the drawings from a bowl.

For our present purpose, we may divide this objective into two parts:

1. To detect and eliminate assignable causes thereby attaining a statistical state.
2. After attaining a statistical state, then to attain the desired characteristics of the frequency distribution of quality.

In more descriptive terms this is like trying to reduce quality variation to a point where the observed values of quality behave as though they were values written on physically similar chips in a bowl and drawn with replacement and thorough mixing. Having attained this condition it is then necessary to find the characteristics of the distribution of numbers in the bowl.<sup>1</sup> Applied to a problem such as securing uniformity in respect to some quality characteristic X of steel, the end result of the two steps would be to establish the frequency distribution of this under an attainable state of statistical control.

Whether or not it is possible to attain the first part of this objective can obviously only be settled in an empirical way. In the first place, what criteria shall be used in the process of control? We must choose from among the indefinitely large number of criteria that would be necessary, as we have seen in the previous section, to characterize a statistical state, one or at least a few simple enough that they can be used in practice. Furthermore the criterion or criteria chosen must take into account two kinds of errors:

- e<sub>1</sub>. Looking for assignable causes when they do not exist.
- e<sub>2</sub>. Overlooking evidence of the existence of assignable causes.

-----  
1. This second step is considered in some detail in Chapter II.



At least one such criterion<sup>1</sup> has been found to work quite satisfactorily as a control operation, in those fields where it has been tried.

One important practical question is: How much data need be taken before we can rest assured that a state of statistical control has been attained? No one can say how much will be required before we begin to get a fairly long run of samples indicating control. However, after preliminary investigations have been made to check on the effect of what the experimentalist in charge considers to be assignable causes and he has, as it were, reached the end of his rope in eliminating these, then not less than 25 samples of four or perhaps more than 250 samples of four are ordinarily required for test in a control chart for averages before one can with much confidence conclude that he has reached a state of control. It will be shown in the next chapter that approximately the same total sample size, 100 to 1000, is required for giving us adequate information about the frequency distribution even when it arises under a state of statistical control.

We shall close this section with an example of what can be done in practice. Fig. 7 shows such a control chart test for averages of 136 successive samples of 10. The quality characteristic is blowing time of a certain kind of fuse. Of course in the preliminary survey assignable causes were in-

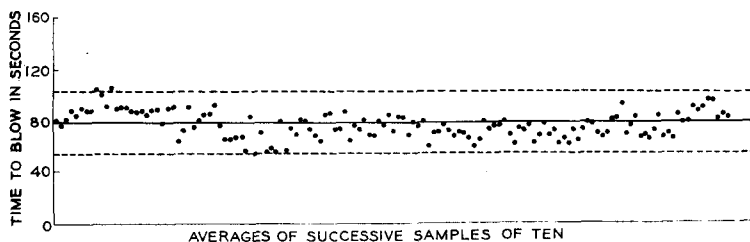


Fig. 7

dicated and removed. This chart is here included as typical evidence that once we attain a condition of control as judged in this way under limitations as to sample size indicated above, this condition seems to continue. The points remain within the control limits almost as well as though they had been obtained from samples from a bowl! That such an apparent state of control can be attained under commercial conditions is all the more impressive when in the

1. See Criterion I in book by Shewhart, W. A., Loc. Cit.

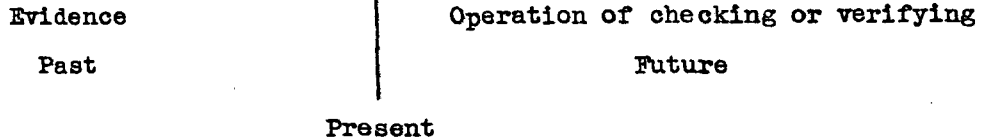
next chapter we compare Fig. 7 with a similar one representing some of the most precise measurements of physical science.

JUDGMENT AS TO STATISTICAL CONTROL

This aspect of statistical control is of vital importance from a practical viewpoint. Let us start with a consideration of the meaning of the statement:

The quality of this product is statistically controlled.

Let us first ask ourselves how we would go about trying to show in the specific case whether this statement is true or false, keeping the level of discussion on a practical plane. What, in other words, would be our operation of verification? There is, however, a second aspect to be considered - do we believe the statement will be found true? Anyone can make such a statement in respect to a given product irrespective of whether or not he has any grounds for believing the statement true. Obviously the operation of checking the statement can only take place at some future time. However, the evidence lies in the past. Schematically we have:



Both the evidence and the operation of verification are important aspects of the practical meaning.

Let us assume that the means of producing the product in question are such that an indefinitely large number of pieces of this kind of product can be made and let us represent as before the quality characteristic under consideration by the symbol X. As already noted, the objects might be any one of the almost countless different kinds of piece parts or materials produced in industry and the quality X might be either chemical or physical. By repeating the process of operation again and again<sup>1</sup>, we would get an indefinite

1. For practical purposes of simplification, it is here tacitly assumed that the process or machine is such as to make but one object at a time. In practice, of course, there is likely to be a whole battery of machines which may turn out more than one piece of product at a time. The treatment here given can be extended to cover this case, but is unnecessarily involved to illustrate the fundamental points here considered.

ly long sequence,

$$X_1, X_2, \dots, X_n, X_{n+1}, \dots, X_{n+i}, \dots \quad (4)$$

Now, of course, one making such a statement might simply refer to the operation of control and the results obtained in the past, but even then the usefulness of this information would only be in the sense of making possible valid operationally verifiable predictions as to the future. One very important class of predictions have to do with predicting the qualities in one or more subgroups in the sequence (4). These subgroups correspond, of course, to what are customarily termed lots. Another important class of predictions have to do with the process or operation of production. For example, it is assumed that if the process - chance causes of variability in product - is in a physical state of statistical control, then one cannot hope to find and remove any more assignable causes. To try to check this statement would involve trying but being unable to find any assignable causes, no matter how long one searched for them.

Now the first point I wish to make is that neither of these operations as stated is experimentally definite in the sense that they can be carried out in a fixed time. In the first case, we need to set up definite criteria to be met by specified limited portions of the sequence. In the second case, we need to state definitely what shall be the nature of the tests to be performed and the requirements to be met by the data thus obtained in order that we shall agree that the assignable causes have been eliminated.

As the next point, let us note the effect of having to choose a specific set of criteria in order to make the meaning definite in a practical as compared with a theoretical sense. We have seen in the discussion of the state of statistical control that there is an indefinitely large number of criteria required to specify the sequence arising from a statistical state. Therefore, any chosen set of criteria must be an incomplete means of checking operationally what we mean by a statistically controlled state.

Let us next consider a little more carefully what is required of the criteria. Obviously they must be in terms of the numbers representing the numerical magnitudes of the qualities of the objects produced by the process in

question. For example, such a sequence should satisfy what I have termed Criterion I, let us say for averages of samples of four drawn with replacement from a normal "bowl-universe". That is, if

$$Y_1, Y_2, \dots, Y_i, \dots, Y_n, Y_{n+1}, \dots, Y_{n+i}, \dots \quad (6)$$

be such a sequence, only approximately three out of a thousand of the averages of ordered samples of four should fall outside the dotted limits. Fig. 8 shows

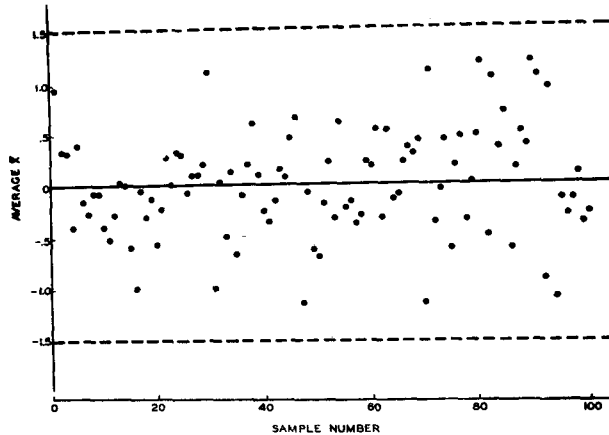


Fig. 8

an experimental check for 100 samples of four. Writing down requirements that a given sequence should satisfy is comparatively simple. The real job is to choose the sequence of X's to represent the process - the statistician might say, choosing how the subsamples to be used in a criterion are to be taken. For example, let us take the case considered above where the engineer or scientist has decided that he is doing the same thing again and again under the same essential conditions. The control chart technique specifies that the sequence for test shall be taken in the order in which the physical things were made. If no requirement were made, so that the sequence might be taken any way whatsoever, then the criterion would be practically worthless. For example, the sequence of measured qualities in this case would perhaps in general be that of the measurements of the objects taken from some storage bin. If the mixing in the process of going to the storage bin is thorough, the chance of detecting lack of control is practically zero. As a graphic illustration of what would be expected to happen in such a case let us look at Fig. 9, a) and b). The

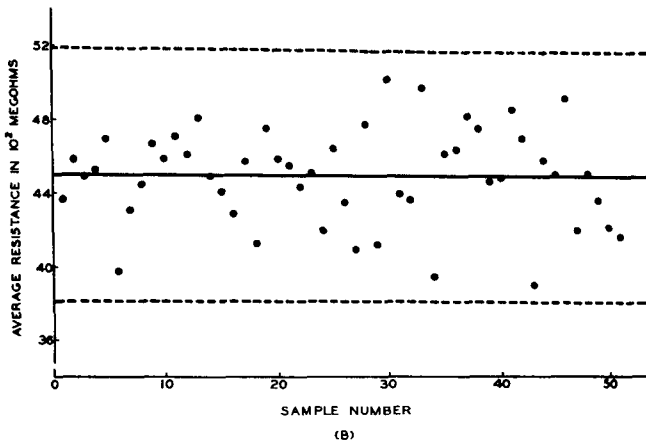
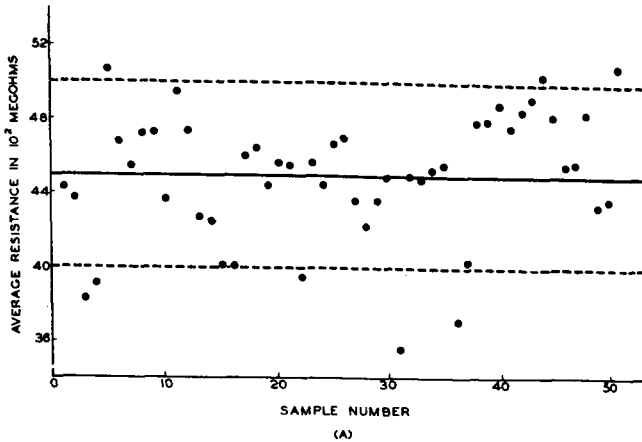


Fig. 9

first of these shows a control chart for 51 averages of four when the sequence was that of the order of production. Fig. 9b shows what happened when the numbers in the first sequence were placed on chips in a bowl and then drawn one at a time. Thus we see as the third point that the problem of giving operationally verifiable meaning to statistical control in practice is two-fold:

- (a) The specification of the way the sequence to be used in any chosen criterion shall be obtained, and
- (b) Choice of criterion (or criteria).

Thus far we have been considering simply the method of checking the prediction implied in the statement that a certain quality characteristic is statistically controlled. Now let us turn our attention to the evidence required for believing such a statement. For example, if this statement were made simply upon the evidence that the scientist or engineer thinks he is operating under the same essential conditions, one could expect that in practice-

ally no instance would the statement be found true as judged by any criterion such as Criterion I. In fact it appears that it is only after the statistical control technique has been applied in practice that the state of control is reached in any appreciable number of cases.

If, on the other hand, one had before himself the record showing how, let us say by some definite control chart technique, assignable causes of variability had been found and eliminated until finally it had been possible to obtain a sequence of at least 1000 objects whose qualities taken in the order of production of the objects satisfies Criterion I for averages of samples of four, he could be quite sure that future product would satisfy any similar criterion. That is to say, the number of times one would find himself in error when making judgments upon the basis of this amount of evidence would be a very small percentage of the total number of trials, whereas it would be practically 100% if judgments were made solely upon the evidence that the conditions had presumably been maintained essentially the same.

#### SIGNIFICANCE OF STATISTICAL CONTROL

Let us first consider the significance of the effect of the study of ways and means of attaining and maintaining statistical control of quality upon statistical methodology. As we have tried to show in the discussion of the state of statistical control, there is a purely formal and mathematical theory of distribution which may be taken as characterizing a purely formal state of statistical control which may or may not be, so far as the formal theory is concerned, descriptive of any state attained in practice. Then there is the concept of a physical state of statistical control, which represents the limit to which we can go in attaining valid predictability and minimum variability. Quality control studies have shown that there is good reason to believe that such a physical state can be attained and when attained it has shown that observables of this state satisfy criteria used in describing the formal state.

In the customary application of statistical theory we assume, of course, that there is a physical state giving samples showing the characteristics of those formally considered in distribution theory. What control studies have shown is that such physical states are indeed rare occurrences, at

least in physics and engineering, and furthermore that they do not, as it were, come into existence without first applying certain operations of control and until comparatively large numbers of preliminary data have been taken in the process of detecting and removing assignable (or findable) causes of variability. In this chapter, we have of course considered the problem of control only from the viewpoint of attaining valid predictability and minimum variability in a measured quality X. In other words, we have neglected the matter of accuracy which will be considered later. We shall then find still more evidence to indicate the need for going through a definite operation of statistical control before applying statistical theory which assumes the existence of a state of statistical control.

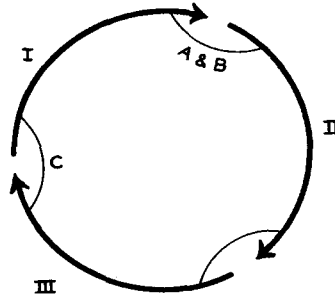
Next let us consider the significance of the study of statistical control from the viewpoint of the control of quality. Let us recall the three steps of control: Specification, production and judgment of quality. On the older concept of an exact science these three steps would be independent. One could specify what he wanted, some one could take this specification as a guide and make the thing, and an inspector or quality judge could measure the thing and see if it met specifications. A beautifully simple picture!

The whole picture is, however, radically different just as soon as we admit that we have only a probable science. Even when we limit ourselves to trying to stay within tolerance limits, it is necessary for economic reasons and for attaining maximum quality assurance in all cases including that where tests are destructive, to introduce the concept of action limits A and B and the aimed-at value C, Fig. 6. But in order to specify C we must first apply the operation of statistical control. In fact the value of C must really come from Step III and after suitable action limits A and B have been established in Step II. But these cannot be set without knowing something about at least the tolerance limits that are specified in Step I. I think it is particularly significant to note that the third step cannot be taken by simply inspecting the quality of the objects as objects but instead they must be inspected as an ordered sequence in relation to the production process and for reasons discussed in the previous section. In fact these steps must go, as it were, in

a circle instead of a straight line as shown schematically in Fig. 10.



OLD



NEW

Fig. 10

#11060  
(Reference loop)

From the viewpoint of specification, it is of interest to note that for the meaning of control to be operationally definite, not only must certain criteria of control be chosen but also the operation of selecting objects whose qualities are to be tested by the criteria must be specified. The choice of criteria to be used as a method of verifying the state of control can be made without reference to a product but the method of specifying the sequence of values to be used in the chosen criteria cannot in general be set down without reference to the results obtained in production. What is still more important, the intent of any such specification implies a certain degree of assurance that the quality of product will be found to satisfy this set of criteria, particularly when the product cannot be given a 100% test. Here again without a knowledge of the results of prior attempts to control quality one cannot specify in a perfectly definite way just how much data are required and just the sequence in which these data shall be used in applying control criteria to give the quality assurance intended by the design specification. For these reasons, it seems necessary that operationally verifiable control requirements and requirements as to how much and how data shall be obtained to provide adequate quality assurance can only be set down in Step III by one having his eye both on the intent of design requirements and upon the accumulated results to date indicating the degree to which a state of statistical control has been



approached. Hence the design specification must be supplemented in Step III by inspection specifications providing adequate data and satisfactory criteria of control for each type of product.

Furthermore, since the running record of past results must play such an important part in judging the degree to which control has been attained, it is necessary for Step III to provide such a continuing record or quality report. The mathematical theory of distribution characterizing the formal and mathematical concept of a state of statistical control constitutes an unlimited storehouse of helpful suggestions from which practical criteria of control must be chosen and the general theory of testing statistical hypothesis must serve as a background in guiding the methods of making a running quality report that will give the maximum service as time goes on.

To attain economic control and maximum quality assurance, statistical theory and techniques must enter every one of the three steps in the control of quality. In this way they make possible a very important potential contribution of mass production to scientific industrial progress. Incidentally we have seen that this potential state of economic control can only be approached as a statistical limit even after the assignable causes of variability have been detected and removed. Control of this kind cannot be reached in a day. It cannot be reached in the production of product in which only a few pieces are manufactured. It can, however, be approached scientifically in a continuing mass production.

4)  $Q$  is sometimes tabulated  
~~if~~  $K \pm \Delta X$

2 Can it be done -

3 How predict in this case  
Remember the cowboy in the  
Texas saloon.

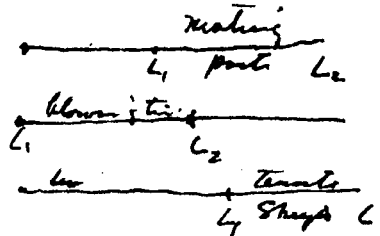
# HOW ESTABLISH LIMITS OF VARIABILITY

## I THE PROBLEM. (PRACTICAL)

a) stated by DA JEWETT.

b) Relation to economic context.

c) Figure on board



d) Example of loss triangle - Standard

$S_1$       $S_2$       $S_3$   
Measurable Iron

## II THE PROBLEM - STATISTICALLY.

a) Find frequency function

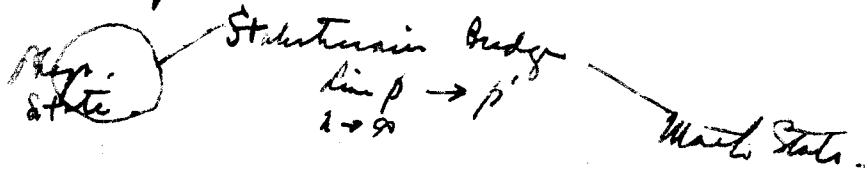
$$dN = f(x, \theta_1, \theta_2, \dots) dx$$

Such a function exists

b) Possible to say,  
 $p \leq p'$

c) Poss a way to improve predictability  
in stability sense.

This implies that we can use the



Illustrate with those of a tie

d) Practical problem (Does  $p'$  exist in this sense)

e) Criterion of control says no

### III How Establish LIMITS: SIMPLEST CASE - THE BOWL

a) Examples

1) Non-verifiable

Board statistics

$S_9$

b) Non-Verifiable Prediction (practice)

Students  $S_9$

1) Deming has designed admirably

c) Verifiable Prediction (tolerance)

Watch out for errors,  
importance of sample size

IV

HOW ESTABLISH LIMITS - PRACTICAL

a) From tabulated values ?

$$x \pm \Delta x$$

Den. 7.871  $\pm$  0.00 =  $\frac{7.871}{c}$   
precision

b) Case of variables in  
remember to write out the  
how look at

c) L C and S.

d) C - True values

e) C in frequency, curve  
within cell is spread fit

f) C certainly was not starting  
estimate, test

## CHAPTER II

### HOW ESTABLISH LIMITS OF VARIABILITY

Thus in many directions the engineer of the future, in my judgment, must of necessity deal with a much more certain and more intimate knowledge of the materials with which he works than we have been wont to deal with in the past. As a result of this more intimate knowledge his structures will be more refined and his factors of safety in many directions are bound to be less because the old elements of uncertainty will have in large measure disappeared.<sup>1</sup>

FRANK B. JEWETT, President  
Bell Telephone Laboratories

Broadly speaking, the problem to be considered here is that of controlling quality of manufactured materials and piece parts so as to provide the engineer of the future with the knowledge necessary in order that he may set his tolerances so as to make the most efficient use of materials and still maintain adequate quality assurance. To begin with, we should clearly differentiate between this problem and that which I have discussed in the literature under the title of economic control of quality of manufactured product.<sup>2</sup> In general, the latter starts with the assumption that tolerance limits on each specified quality characteristic have already been set. This means that for any given characteristic  $X$  there are certain limits  $X_1$  and  $X_2$  such that the quality  $X$  of each piece of the product is supposed to lie within this range. We shall speak of a piece of product with a quality  $X$  meeting this requirement as conforming in respect to the specification of  $X$  and likewise, a piece of product not meeting this requirement shall be spoken of as non-conforming. Now, of course, non-conformance usually implies rejection or at least modification of the part in question. Hence it is of economic importance to minimize the number of pieces non-conforming. Here then we have the problem of economic control within specified tolerance limits on the piece. There is, of course, another aspect to this problem of manufacturing to tolerance limits on each piece when the inspection test is destructive and where it is necessary to have a very high degree of assurance that the quality of the piece conforms to tolerance requirements even though it cannot be tested.

-----

1. "Problems of the Engineer", Science, Vol. 75, No. 1940, March 4, 1932.

2. Shewhart, W.A. loc. cit.

It was for the purpose of effecting economies in production and of attaining maximum quality assurance at a given cost that the quality control chart technique involving the introduction of the concept of two action or control limits and an expected value was introduced in 1924.

The problem with which we shall be concerned in this chapter is, however, that of choosing the tolerance limits themselves which in the application of control chart theory are assumed to be given by the design specification. Again considering the three steps in control, specification, production and judgment of quality, the problem of setting tolerance limits comes under step I. However, for reasons emphasized in the previous chapter, in order for these limits to be practically attainable, they must be set with an eye on what can be attained in commercial production. Quite naturally engineering practice has always taken into account the necessity of stating tolerances that are thus consistent with what it is believed can be done in practice. Furthermore this was done with marked success for many years before any one perhaps even thought about applying statistical techniques in the process of control. The object in this chapter is to consider briefly some of the fundamental aspects of the problem of setting tolerances and to indicate some of the potential contributions as well as inherent limitations of the application of statistical theory to its solution.

To fix the problem to be considered, let us confine our attention to a single quality characteristic  $X$ . Three typical cases arise in practice as illustrated schematically in Fig. 11. Let  $p$  represent the probability of

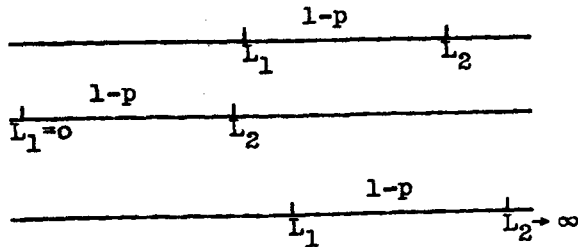


Fig. 11

a value of  $X$  falling outside of tolerance limits  $L_1$  and  $L_2$ . The problem may then be thought of as establishing tolerance limits in such a way that

$$p \leq p' \quad (7)$$

where  $p'$  is some given value. The first of the cases shown in Fig. 11 corresponds to the requirements for mating parts. In addition to the requirement (7) we may also wish to require that

$$R = (L_2 - L_1) \text{ must be a minimum.}$$

As an illustration of a tolerance range with only an upper limit to be set we might take the requirement that the blowing time of a fuse shall not be greater than  $L_2$  seconds. Likewise there are many examples of a requirement in the third form such as, that the tensile strength of the steel strand shall not be less than  $L_1$ . In either of the last two cases there is implied some value of  $p'$ . For example, in setting safety factors it is desirable to make  $p'=0$ .

Our object in this chapter may now be a little more definitely defined as that of trying to determine at least some of the potential contributions and inherent limitations of the application of statistical theory in the establishment of the tolerance limits  $L_1$  and  $L_2$  in each of the three cases.

#### PROBLEM FROM VIEWPOINT OF STATISTICS

The problem as thus stated appears to have many of the earmarks of the so-called statistical problem of estimation - a subject to which a great amount of attention has been given by theoretical statisticians and students of error theory. Hence one might expect that all the engineer needs to do in order to improve his technique in setting tolerances is to become acquainted with the available theory of estimation. We shall find, however, that such an expectation is not justified, but in making this statement, we are getting ahead of our story.

Let us suppose that we wish to use malleable iron in some design in which we are interested and therefore want to set tolerance limits on tensile strength. We naturally turn to the engineering literature for data obtained under practical conditions to be used as a basis for estimating our tolerances. In the report of a recent symposium on malleable iron we find the results of 5000 tensile strength tests on malleable iron test bars made by Enrique Touceda

-----  
1. Symposium on Malleable Iron Castings, published in Proc. Amer. Soc. Testing Mat., Vol. 31, 1931, pp. 317-434.



for the Malleable Iron Research Institute, the bars having been taken over the period from May to November, 1930, from "random" heats of the companies comprising the membership of the Institute. These data are presented in Table 1.

Tensile Strength of 5000 Malleable Iron Bars

<u>Range of Values lbs. per sq. in.</u>	<u>Observed Distribution</u>	<u>Normal Law Distribution</u>	<u>Difference</u>
Under 45,000	0	0	0
45,000 - 45,999	1	0	1
46,000 - 46,999	2	1	1
47,000 - 47,999	3	5	2
48,000 - 48,999	8	22	14
49,000 - 49,999	23	77	54
50,000 - 50,999	289	210	79
51,000 - 51,999	472	448	24
52,000 - 52,999	739	744	5
53,000 - 53,999	927	963	36
54,000 - 54,999	967	970	3
55,000 - 55,999	758	762	4
56,000 - 56,999	481	466	15
57,000 - 57,999	230	222	8
58,000 - 58,999	72	82	10
59,000 - 59,999	19	24	5
Over 60,000	9	5	4

Table 1

Here we have a very respectable looking uni-modal frequency distribution. In fact when one graduates this distribution with the normal law (column 3 of Table 1) and plots the results, he gets Fig. 12. It may seem quite reasonable

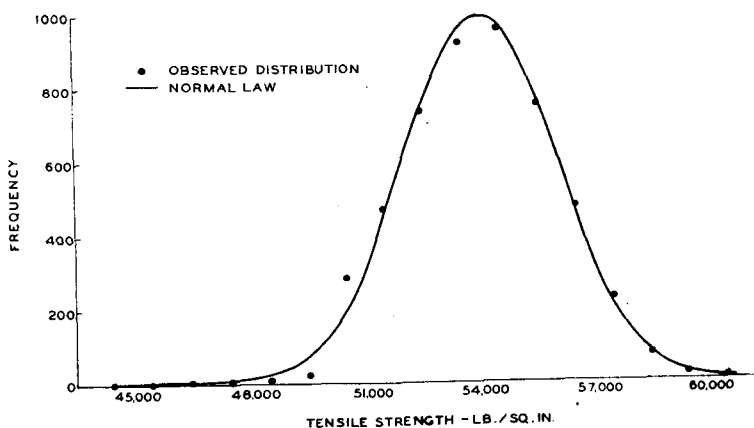


Fig. 12

therefore to think of this distribution of 5000 observed values as a "random" sample of a "hypothetical universe" which the production process is capable of giving if allowed to function indefinitely. Since the closeness of fit between the theoretical graduation and the observed distribution as measured

by a  $\chi^2 = 90.23$  is not very good, the theoretical statistician may argue that the hypothetical universe is not normal. For our present purpose, however, we are not concerned with the functional form of the universe but merely with the assumption that the universe exists. If it exists in this customary statistical sense, then it would appear that the problem of setting tolerances reduces to a statistical problem of estimation. Our problem from this angle is therefore two-fold: a) to examine the justification, in general, of the assumption that a statistical universe exists and b) to consider the technique of setting tolerances when the assumption is justified.

There is, however, another aspect of the subject of setting tolerances which we must investigate and we can perhaps approach this best through an illustration of the way the problem arises. Suppose we wish to make use of pure iron in some way that requires us to set tolerances on its density. Accordingly we turn to an authoritative table<sup>1</sup> of physical properties and find the density given as  $(7.871 \pm 0.002)$  gms/cm<sup>3</sup>. This example is taken as typical of the case where the available information upon which to base estimates of tolerances is given in the form

$$X \pm \Delta X.$$

Among other things, we must examine the meaning of this range as customarily given and see what relation it bears, if any, to the tolerance range. If we turn to the literature of modern statistics, we find much emphasis placed upon the fact that ranges of type (9) can be established upon the basis of modern small samples that are just as valid as are those based upon large samples. Now, of course, a tolerance range can also be put into this form in so far as its numerical aspects are concerned. Hence the engineer rightly wants to know if the statisticians have found a royal road in the sense that it can be based on estimates from small samples. We shall find an abundance of confusion on this point even in the literature of statistics. The fact that the meaning of the range which is valid in the sense of so-called modern small sample theory turns out to be different from the meaning of the tolerance

-----  
1. Physical Constants of Pure Metals, The National Physical Laboratory, England, 1936.

range should be of considerable interest to statisticians as such as well as to engineers.

From the viewpoint of presentation much is to be gained by starting with the simplest case where we know that the sample of data with which we start was drawn one at a time with replacement from an experimentally<sup>1</sup> normal universe.

HOW ESTABLISH TOLERANCE LIMITS - SIMPLEST CASE

Let us assume that we have drawn a sample of  $n$  values  $X_1, X_2, \dots, X_1, \dots, X_n$  from a normal universe. The problem to be considered first is that of setting up a tolerance range  $X = L_1$  to  $X = L_2$  that will include let us say  $(1-p')N = .5N$  of future drawings from the bowl. An engineer might wonder why we choose  $(1-p') = .5$  whereas in practice it is more likely to be less than .01. We choose this value of  $p'$  because several books in science and in error theory seem to tell one just how to establish  $L_1$  and  $L_2$  in this case. For example, one outstanding treatise of 1937 on a particular branch of physics has an appendix discussing accuracy and precision. The authors give eleven measurements of a length and calculate the arithmetic mean  $\bar{X}$  of the sample and the probable error  $e$  of a single observation in accord with classic error theory. They then state in effect that if a further series of  $n$  measurements are made under the same conditions, it is an even chance that the mean of the second series of  $n$  observations will differ from the mean of the first series of  $11$  measurements by more or less than  $\frac{e}{\sqrt{n}}$ . This certainly looks to the uninitiated like just the solution to his problem assuming that he wants a limiting range corresponding to a probability of  $1/2$ . Of course, the implication is that ranges for any probability could be set up in an analogous manner with proper allowance for the magnitude of the probability  $p'$  that is chosen.

Sometimes seeing is more convincing than reading an argument. Hence let us see what might happen to one who followed such a rule. For this purpose, I drew from a normal universe in a bowl the sample of eleven measurements shown in Table 2. The average  $\bar{X}$  and probable error are .0091 and .3226 respectively. Now let us set up tolerance limits for a probability of  $1/2$

-----  
1. See, for example, page 165, Table 22, of my Economic Control of Quality of Manufactured Product, for such a distribution.

and sample size of  $n = 4$ . According to the previous paragraph, such limits

.5	.1	-.3	-.9	.1	-.1
-.1	.3	-.6	.7	.4	

Table 2

would be  $.0091 \pm .1613$ . According to the authors of the text under consideration, we should expect to find fifty per cent of samples of four to have averages lying within this range. Well, let us take 100 samples of four and see if such a prediction is valid. Fig. 13 shows the results of one such test. We were led to expect 50% within the limits  $.0091 \pm .1613$  shown by dotted lines.

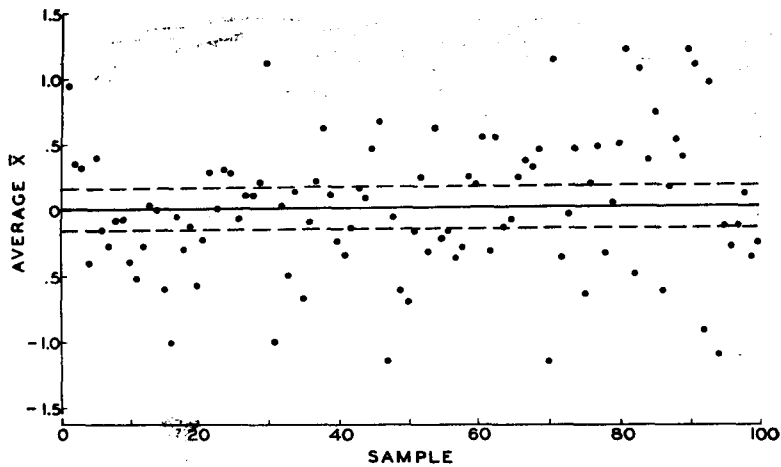


Fig. 13

We find 27%! The prediction of 50% within limits wasn't valid!

I am reminded of the old jingle: When a doctor makes a mistake, he buries it; when a judge makes a mistake, it becomes the law, and so on. I would add: When scientists such as the authors of the book referred to and many others make such a mistake, no one seems to challenge it; but when an industrial statistician makes such a mistake, woe unto him for he is sure to be found out and get into trouble. For example, in establishing tolerances one can rest assured that he will hear about it if appreciably more than the expected percentage are found outside of limits and furthermore he can rest assured in most instances that any such tendency will be discovered because hundreds, thousands and sometimes even millions of pieces of product are made per month.

16473  
17800

It would, of course, be unfair for the engineer to judge the contribution of statistical theory from experiences such as the one just considered which fails to take account of the important recent contributions to statistical theory which overcome some inherent limitations of the older error theory. I have in mind, in particular, the work beginning with "Student's" publication<sup>1</sup> in 1908 of tables for the probability  $P_z$  that the mean of a small sample of  $n$ , drawn at random from a population following the normal law will not exceed in the algebraic sense the mean  $\bar{X}'$  of that population by more than  $z$  times the standard deviation of the sample. Let us therefore see if this fundamental contribution to the theory of the error of the mean helps us to solve the practical problem of establishing tolerance limits for the ideal case of a normal bowl universe. First we shall try to determine just what the theory enables one to predict in a valid way. Interpreted in an operationally verifiable way this theory simply means, among other things, that given a normal universe whose average  $\bar{X}'$  and standard deviation  $\sigma'$  are unknown, we can make the valid prediction that if we draw a series of  $N$  such samples of size  $n$  from the given population and calculate ranges  $\bar{X}_i \pm z\sigma_i$  where  $\bar{X}_i$  and  $\sigma_i$  are the average and standard deviation of the  $i$ th sample of size  $n$ , then  $P_z N$  of these ranges may be expected to include the average  $\bar{X}'$  of the universe. If the universe is an experimental one where the theoretically true value  $\bar{X}'$  can be obtained, then such a prediction can be checked.

For example, Fig. 14 shows a series of 100 such ranges for  $n = 4$ ; 40 for  $N = 100$ ; and 4 for  $n = 1000$ , where  $P_z = .50$ . The theoretical value  $\bar{X}'$  in this case was zero and is shown by the heavy central line. The figure shows that the percentages of ranges including zero are 51, 45, and 50, the expected value  $P_z$  being 50. This constitutes an excellent check between theory and experiment for different size samples.

Thus we see how it is possible with the aid of "Student's" theory, to make predictions about ranges in the sense of Fig. 14 that are just as valid for small samples as for large. It enables one to make valid predictions about a series of ranges of the form

---

1. "The Probable Error of the Mean", Biometrika, Vol. VI, Part 1, 1908, pp. 1-25.

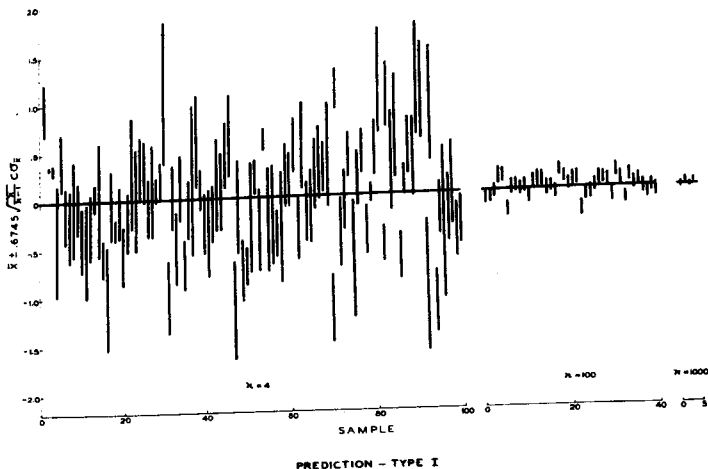


Fig. 14

$$\bar{X}_1 \pm z\sigma_1; \bar{X}_2 \pm z\sigma_2; \dots \bar{X}_i \pm z\sigma_i, \dots \bar{X}_N \pm z\sigma_N; \dots$$

calculated from a sequence of samples of size  $n$  drawn from a normal universe, where  $\bar{X}_i$  and  $\sigma_i$  are the observed average and standard deviation respectively of the  $i$ th sample.

Now let us consider the problem of establishing a tolerance range for the normal bowl used in getting the data presented in Fig. 14. As before, let us assume that we do not know the parameters of the normal distribution in the bowl. Our problem is to set up a range  $X = L_1$  to  $X = L_2$  such that the probability of drawing a value  $X$  from the bowl that will lie within this range is some previously specified value  $p'$ . As a special case let us take  $p' = .5$ . It is of course, assumed that the only way we can find out anything about the normal universe in the bowl is through drawings with replacement. This hypothetical case corresponds to a practical case where it is known that the production process constitutes a normal statistical state but the parameters are unknown.

Obviously the starting point is to draw a sample of  $n$  values of  $X$  from the bowl. To make the problem specific let us assume that we draw the following sample of four

1.7, 0.2, 1.4, 0.5

How shall we set up the tolerance limits  $L_1$  and  $L_2$ ?

I think it will be generally and perhaps unanimously agreed among statisticians that our best estimate of such a range can be put in the general

form  $\bar{X} \pm t\sigma$ . It is obvious, however, that no matter what rule is adopted for computing such a range, the range computed will seldom correspond to a probability  $p' = .5$ . It is also obvious that the problem of establishing a valid tolerance range is fundamentally different from the problem solved by "Student". "Student's" theory tells how to make valid predictions about how many times a series of ranges may be expected to include a theoretically true value, whereas, in order to establish a valid tolerance range, we must be able to make a valid prediction about how many times future observed values may be expected to fall within a range computed from a sample of  $n$ .

The fundamental difference between what we might term the "Student" type of range and the tolerance range is so important that we are perhaps justified in examining the nature of this difference a little more carefully. For the special case where  $P_z = .5$ , the "Student" type range is sometimes called the probable error or the equi-bet range. In this sense, this range is likely to be confused with the probable error range of the classic theory of errors which for averages of samples of size  $n$  would be  $\bar{X}' \pm \frac{\sigma'}{\sqrt{n}}$  where  $\bar{X}'$  and  $\sigma'$  are the true average and standard deviation respectively of the universe. If we knew this range, we could make the valid prediction that 50% of future averages of samples of size  $n$  would be found to fall within this range. In fact, if we knew  $\bar{X}'$  and  $\sigma'$  we could set up a valid tolerance range for any value of  $p'$  for any experimentally normal bowl universe. The probable error range in the "Student" sense is entirely different because the probability  $P_z = .5$  refers not to future averages of  $n$  lying within a range determined from the sample but to a series of ranges determined from samples of size  $n$  including the true value  $\bar{X}'$  of the universe.

Now let us see what we must do in order to set up a tolerance range for which we can make a prediction which is valid within limits that are practical. For this purpose let us choose  $p' = .9973$  because this is roughly of the magnitude customarily desired in engineering practice. Of course if we knew  $\bar{X}'$  and  $\sigma'$ , the desired range would be  $\bar{X}' \pm 3\sigma'$ . Let us see what happens if we take as a range  $\bar{X} \pm 3\sqrt{\frac{n}{n-1}}\sigma$ . Fig. 15 shows 100 such ranges

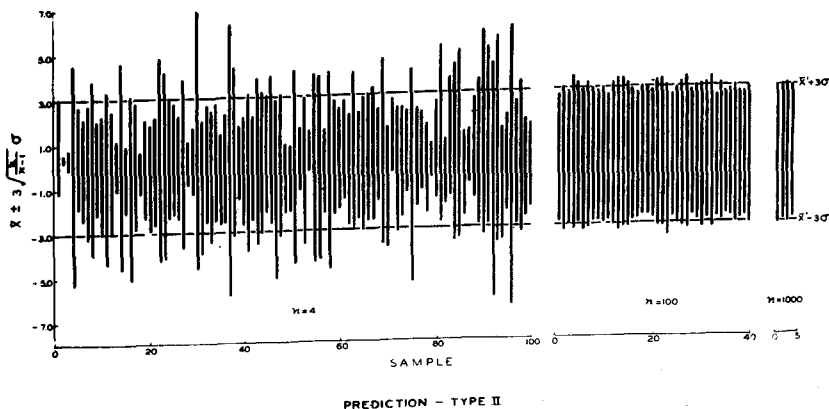


Fig. 15

for as many samples of 4 drawn from an experimental universe; 40 ranges for 40 samples of 100 and 4 ranges for 4 samples of 1000. The dotted limits are  $\bar{X} \pm 2\sigma'$  or the true values as one would say. On this figure hangs a tale of great practical importance.

If one were to go through life setting 99+% tolerances by the method indicated above on the basis of samples of 4, even when drawn from a pedigreed normal universe, he would sometimes get a range including very small percentages. For example, the second range in Fig. 5 includes only 12% instead of the ideal 99+%. Furthermore he would seldom get ranges approximating very closely to symmetry. Even on the average the ranges thus set up would not include the ideal 99+% but something less. The average observed for 1000 such ranges is 93%.

Of course, it is theoretically possible to choose a value for  $t$  that will cut out on the average the ideal value 99+% but in that case the errors of the separate ranges would be increased on the average. There are many details of interest that might be considered but for our present purpose it is sufficient to call attention to the fact that experimental ranges have a tendency to hug closer to the ideal limits the larger the sample size used in computing the limits.

This fact is of great practical importance. It shows that if we wish to reduce the chance of making an error of a given size in estimating the probability associated with chosen tolerance limits, there is no royal

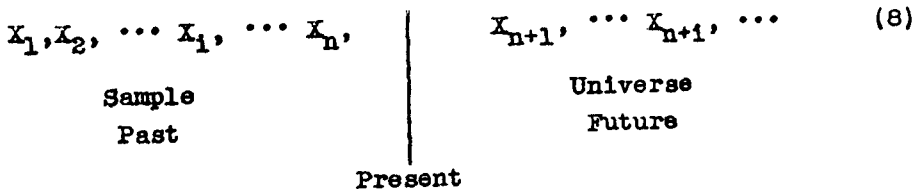


"small sample road" of doing this. Even under the simple conditions here assumed, we can improve our estimate only by increasing the sample size  $n$ . Certainly even under the ideal conditions here assumed, viz., normal bowl universe, one would not be likely to be satisfied with a sample size less than 1000 and certainly not less than 100 if he were trying to set tolerance limits that would insure the most efficient use of engineering materials. That is to say, even if the properties of materials and manufactured product were in a state of statistical control to begin with, it would still be necessary in order to acquire the more certain and intimate knowledge required for setting the most efficient tolerances, to have a sample of at least 100 and more likely 1000 or more.

Another very important point for us to note is that there is no way under the shining sun to form an estimate of the errors that might be made in adopting a tolerance range of the form  $\bar{X} \pm t\sigma$  unless we know the sample size  $n$  from which it was computed.

HOW ESTABLISH TOLERANCE LIMITS - PRACTICAL CASE

Thus far we have considered the method of establishing tolerance limits, assuming that the world were a bowl of chips. Under these conditions we can only increase our knowledge upon which to base our estimates of tolerance limits through the process of taking more data - a larger sample as it were. The problem is purely statistical in character in the sense that any sample of  $n$  observed values may be considered as a sample of an indefinitely long sequence of numbers satisfying the requirement that they come from a statistical state of control. Schematically we have:



How to set up tolerance limits  $L_1$  and  $L_2$  upon the basis of the sample, and how to determine the errors that may be expected depending upon the sample size  $n$  is a problem to which the mathematical statistician can contribute more than

any one else, assuming, of course, that the physical state of statistical control represented by the bowl can be characterized by the mathematics of distribution theory characterizing the concept of a statistical state of control in the sense already discussed in Chapter I. In fact, there is presumably nothing that an experimental scientist can tell a statistician about how the  $n$  numbers arose that contributes to the work of the statistician beyond the statement that they were drawn from a bowl. Thus we see that in so far as the state of statistical control represents the limit to which one can hope to go in attaining uniformity of quality of product, the problem of setting the most efficient tolerances reduces in the end to a purely statistical problem.

Now let us ask how often in the practical field is one justified in assuming upon the basis of a small sample of data that the conditions have been maintained essentially the same in the sense that one would be justified in making predictions such as indicated above? A mathematician obviously cannot answer this question. We must appeal to experience for an answer, but in analyzing and interpreting the experience the statistician and scientist must cooperate in a way which we shall now consider briefly.

To make our problem a little more specific, let us assume that we are given a set of sixteen measurements, Table 3, of a physical quantity and that we wish to set up tolerance limits for such measurements. What should be

6.683	6.681	6.676	6.678	6.679	6.672	6.661	6.661
6.667	6.687	6.664	6.678	6.671	6.675	6.672	6.674

Table 3

our first step? Shall we call in a statistician and ask him to proceed as if the sample had been drawn from a bowl, or shall we first consult the scientist who took the measurements and ask him some definite questions and, if so, what shall we ask him?

The procedure which some engineers, scientists and statisticians follow is to make a distinction between the observations of the highly skilled and technically trained research worker and the kind of data with which an

engineer must often work. There is a tendency to place a kind of halo around the data of research as though they represent the very last word on the matter. In fact the physicist or chemist often disdains to take more than five or ten observations as though there were nothing to gain by taking more - that is he may appear to base his conclusions on small samples. In contrast the engineer usually wants a lot of data. Now of course, it is true that if all assignable causes of variability have been removed or, in other words, if the small (or large) sample of the research man is such as one might draw from a bowl, then the statistician should be best able to make valid predictions upon the basis of this sample.

As noted in Chapter I, scientists have a habit of saying when they think they have done an excellent job measuring some physical constant or property that all the measurements were made under the "same essential conditions". The statistician as a rule not knowing any too much science and the scientist not knowing any too much statistics, the two have often gotten together and agreed, as it were, that the phrase "same essential conditions" can be taken as a pass word between the two groups.

From this viewpoint then, one might conclude that it would be sufficient for us in considering the data of Table 3 to ask if they had a good research pedigree. Then, if they were approved by a man of outstanding authority in the field from which the data came, we might be tempted to turn the problem over to the statistician to tell us what he could upon the assumption that the 16 data constituted a sample from a bowl of chips.

The same people who might agree to this procedure for the case of data of research would likely question the application of the same procedure to the problem of setting tolerance limits solely upon the basis of considering the 5000 observed values of tensile strength of test bars of malleable iron castings, Table 1, as a sample from a bowl universe. If you were to give them an argument, they would likely point to the fact that in the same reference from which these 5000 data were taken, there were summary figures given in the form of means, maxima and minima for large samples of similar

tests from seventeen different sources. These data are presented in Table 4.

<u>Investigator</u>	<u>Tensile Strength</u> <u>lb. per sq. in.</u>		
	<u>Maximum</u>	<u>Minimum</u>	<u>Average</u>
No. 1	59 000	45 000	54 000
No. 2	58 500	53 000	56 250
No. 3	56 880	50 000	52 460
No. 4	55 850	47 850	52 890
No. 5	62 140	54 400	57 920
No. 6	62 860	52 150	56 350
No. 7	56 000	50 000	53 000
No. 8	58 000	50 000	55 000
No. 9	61 300	49 000	55 000
No. 10	59 800	50 000	53 970
No. 11	60 600	46 600	52 670
No. 12	58 000	50 000	53 000
No. 13	62 000	51 000	53 000
No. 14	56 640	45 500	51 170
No. 15	61 500	45 000	53 710
No. 16	58 000	50 500	55 500
No. 17	56 160	50 480	<u>52 830</u>
Average Tensile Strength .....			54 040

Table 4

The total number of tests summarized in Table 4 is upwards of 20,000. In spite of the fact that the averages of the two sets 54,040 lb. per sq. in., and 54,030 lb. per sq. in., as well as the resultant ranges, are not very different, I think that both statisticians and engineers would agree that it is pretty likely that the chance cause system behind the 5000 test values is not free from assignable causes and hence that these values cannot be considered as a sample from a bowl.

Now, let us look at some of the series of data taken in pure science to see if they appear as if they had been drawn from a bowl. Let us in fact look at the scientists' record in measuring three of the seven fundamental constants of physical science, namely the velocity of light  $C$ , gravitational constant  $G$ , and Planck's constant  $h$ . Certainly such measurements may be expected to be among the elite of all physical measurements. Fig. 16 shows the

fluctuation in the accepted values over the past years<sup>1</sup>. The three ordinate

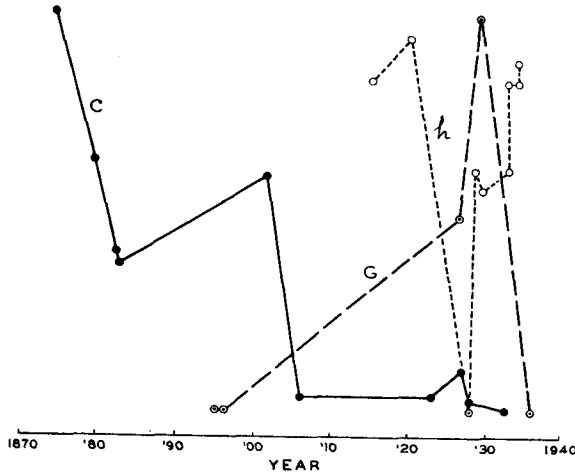


Fig. 16

scales are not shown as the object here is simply to indicate in a readily comparable way the variations in time in each of the three sets of measurements. On the face of the evidence here presented it might be argued that, for the velocity of light  $c$ , the measurements seem to be approaching asymptotically to some fixed value. This type of argument has, in fact, been advanced by Bavink<sup>2</sup> as indicating the more or less ordered way we approach perfect knowledge in physics. The other two curves, however, constitute quite a contrast. Each ends at approximately the level it began. Physicists are pretty generally in agreement that in each of the three cases the observed range of variation is indicative of the existence of "constant" errors. If, however, we examine some of the points at the extremes of the ranges shown, we still find evidence consistent with the hypothesis that there are assignable causes of variability present.

1. "The Velocity of Light", R. T. Birge, Nature, Vol. 134, page 771, Nov. 17, 1934. "On the Values of the Fundamental Atomic Constants", Sten von Friesen, Proc. Royal Society London, A, No. 902, Vol. 160, pp. 424-440, June, 1937. The values of  $G$  for 1895 and 1896 are taken from the article on Gravitation in the Eleventh Edition of the Encyclopedia Britannica. These are the values which the author of the article, J. H. Poynting, thought most likely to be correct at that time (1910). The 1927 and 1930 values are those given in the Smithsonian Physical Tables, 1933, while the 1936 value is obtained from "Fundamental Physical Constants", by W. N. Bond, Phil. Mag., Ser. 7, Vol. XXII, p. 624, October, 1936.

2. Bavink, Bernhard, The Anatomy of Science, G. Bell and Sons, Ltd., London, 1932.

For example, let us take the case of the last determination of the velocity of light shown on the chart<sup>1</sup>. Here is a case where the total number of repetitive observations is large - 2885 in fact. If these readings could be considered a random sample from a normal bowl of chips with an average equal to the true velocity of light, we could be pretty sure that 99.7% of all future observations would fall within the range of the average 299,774 plus 3 times the observed standard deviation. But as is almost always the case where large samples are available, the sample does not give much evidence of having come from a normal universe. Fig. 17 compares the observed distribution with the fitted normal curve. The  $\chi^2$  test tells us that the probability of getting a deviation from normality (as measured by  $\chi^2$ , of course) as large as or larger than that observed, is too small to be read from the tables of  $\chi^2$ . Hence in

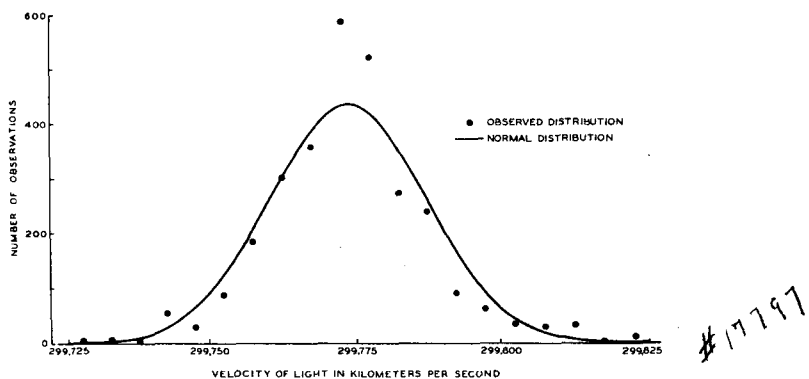


Fig. 17

this case, if one wished to set up valid tolerance limits on future observations of the velocity of light, he would be unwise to use a rule based upon the assumption of normality.

But - and this is the most important question - are we justified in believing that these data constitute a random sample from a constant system of chance causes of variation? Suppose we let the data speak for themselves when successive groups are plotted in the form of a control chart<sup>2</sup>, Fig. 18. The chance of an average going outside the dotted limits if the samples come from

1. Michelson, A.A., Pease, F.G., and Pearson F., "Measurement of the Velocity of Light in a Partial Vacuum", Astrophysical Journal, Vol. 82, 1935, pp. 26-61.
2. Criterion I as described on p. 309 of Economic Control of Quality of Manufactured Product is here used.

a constant cause system is of the order of magnitude of .003. Hence the fact that four points in a total of forty-six fall outside is indeed not very likely

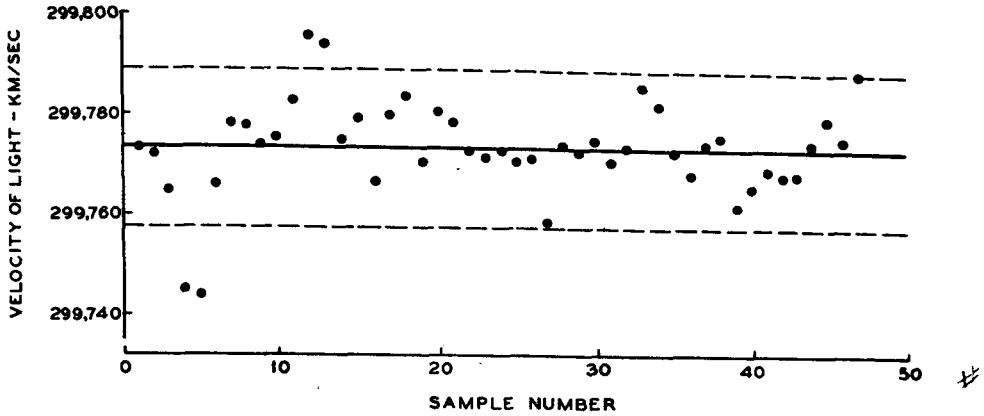


Fig. 18

on the assumption of constancy of the cause system of variation.

What is the practical significance of this fact for our present story? It is simply this: Whereas there is safety in numbers in setting tolerance limits on the basis of a sample from a bowl, that same degree of safety does not exist when samples are not so drawn. My own experience has been that when a set of past data behave as they do in Fig. 18 it never pays to pin one's faith in numbers alone.

Now let us look at another point - this time the maximum point shown on the G curve. This value  $6.670 \times 10^{-8} \text{ cm}^3 \text{ g}^{-1} \text{ sec}^{-2}$  is that given by Heyl<sup>1</sup>. It was derived from the three sets of measurements shown in Table 5 corresponding to experiments using platinum, gold and glass spheres. The value given by

<u>Gold</u>	<u>Platinum</u>	<u>Glass</u>
6.683	6.661	6.678
6.681	6.661	6.671
6.676	6.667	6.675
6.678	6.667	6.672
6.679	6.664	6.674
6.672		

Table 5 - Values of G in units of  $10^{-8} \text{ cm}^3 \text{ g}^{-1} \text{ sec}^{-2}$

Heyl is obtained by weighting the results obtained with gold spheres by one third - and the others by unity. These results are shown graphically in Fig. 19. Certainly we need no refined tests to convince ourselves that such

1. Heyl, Paul R., Journal of Research of Bureau of Standards, Vol. 5, 1930, pp. 1243-1290.

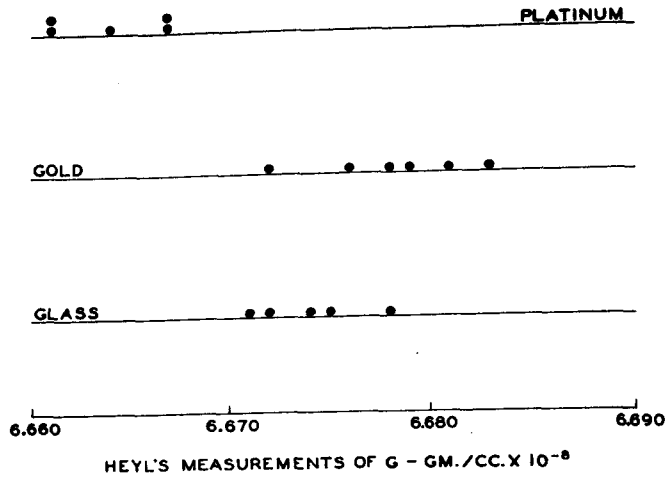


Fig. 19

a set of data is very unlikely upon the assumption of a constant system of chance causes being the source of the observed variability. For example, Heyl says: "The different results obtained with the various materials used for the small masses are yet to be explained, but evidence is given that this difference is not to be ascribed to the nature of the material". The point I wish to stress is that here again we have a sample of measurements among the most elite of pure science that do not seem to behave like drawings from a bowl of chips.

Now let us return to the question under discussion. Given the data of Table 3, shall we first call in a statistician or shall we seek out the scientist who is an authority in the field from which the data came? It would seem that evidence for lack of control of the measurements of the constants  $c$ ,  $C$ , and  $h$  might serve to shake our faith in taking the scientist's judgment that the conditions have been maintained essentially the same as a satisfactory basis for turning the data over to a statistician to be considered simply as a sample from a bowl. In fact, as the reader may have already noted, the sixteen measurements of Table 3 are the same as those of Table 5, except for a constant multiplier! In the light of such experience in the investigation of available data in the field of physical science and from my experience in the study of samples of measurements of quality in engineering, I feel that before one sh-



turn over any sample of data to the statistician he should first ask the scientist or engineer for evidence of statistical control. The statistician's work begins, as it were, after the scientist tells him whether or not the sample has arisen under statistically controlled conditions. The case is something like the old story about the Irishman Pat who had been in this country only a few months and in the meantime had located a job as a hod-carrier when his friend Mike arrived. "Pat" says Mike "And what are you doing"? To which Pat answered: "Sure an' I have an easy job. I carry the bricks up four flights of stairs and the man up there does all the work". In much the same sense the scientist must carry his data through several control criteria before handing them over to the statistician.

But there still remains the question as to how to set tolerance limits even when the chance cause system is not in a state of control. Certainly the engineer must set tolerances and the scientist must form estimates of ranges within which measurements of physical constants and properties may be expected to lie even when conditions do not give evidence of control. All that we have attempted to show thus far is that this isn't a problem to be turned over to the statistician to solve on the assumption that the available data can be treated as a sample from a bowl.

#### HOW ESTABLISH TOLERANCE LIMITS - FURTHER CONSIDERATIONS

As a starting point for what follows, we need to look somewhat more critically than we have done thus far to the requirements that tolerance limits must meet in the process of mass production of interchangeable parts. Thus far we have spoken only of tolerances expressed in terms of measurements of some quality characteristic  $X$ , it being tacitly assumed, as it were, that one thing or piece part may be substituted for another if the measurements of both parts fall within the prescribed tolerance limits. Obviously, however, this condition may be satisfied and one part still not be interchangeable with another for the simple reason, as we say, that the measurements may be in error. Strictly speaking, therefore, we need to differentiate between the customarily accepted concept of true value  $X'$  of a physical quality and a measurement  $X$  of

this true value. Thus if we have two objects  $O_1$  and  $O_2$ , we customarily assume that the true values  $X'_1$  and  $X'_2$  of the quality characteristic  $X'$  must lie within some tolerance range

$$X' = L_1 \text{ to } X' = L_2 \tag{9}$$

in order for the objects to be considered as interchangeable. Likewise the physical state of statistical control would be expressible in terms of a sequence of numbers representing true values, -

$$X'_1, X'_2, \dots, X'_i, \dots, X'_n, X'_{n+1}, \dots, X'_{n+1}, \dots \tag{10}$$

Let us look a little more closely at the concept of a true value  $X'$ . How is one to know whether or not the true value of the quality characteristic lies within a given range? If one cannot knowingly discover the true value then of what practical use is the concept? In answer we shall see that it gives rise to a class of operationally verifiable criteria which are quite distinct in certain characteristics from those used in the operation of statistical control.

To begin with, let us note that corresponding to every concept of a measurable quality  $X'$  such for example as length, there are usually at least several ways of measuring the quality in question. For example, some of the ways of measuring length are: a) use of ordinary rule, b) micrometer, c) traveling microscope, d) triangulation and so on indefinitely. Let us represent the measurement by any such set of methods by the symbols  $X_1, X_2, \dots, X_i, \dots$ . Presumably each method can be repeated again and again at will so that corresponding to any true value  $X'$  there are potentially as many infinite sequences as there are known methods of measuring. Schematically we have:

$$X' \xrightarrow{0} \left\{ \begin{array}{l} X_{11}, X_{12}, \dots, X_{1i}, \dots, X_{1n}, X_{1n+1}, \dots, X_{1n+i}, \dots \\ X_{21}, X_{22}, \dots, X_{2i}, \dots, X_{2n}, X_{2n+1}, \dots, X_{2n+i}, \dots \\ \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \\ X_{i1}, X_{i2}, \dots, X_{ii}, \dots, X_{in}, X_{in+1}, \dots, X_{in+i}, \dots \\ \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \end{array} \right. \tag{11}$$

Where the symbol  $\xrightarrow{0}$  stands for the operational meaning of  $X'$ .

Customarily we tacitly assume that in order for the set of sequences to constitute a measure of the true value  $X'$ , it is necessary that each sequence represent a statistically controlled condition and that the statistical limits of the sum of the first  $n$  terms of any one of the sequences be equal to the statistical limit for each of the others or that

$$\bar{X}_1 = \bar{X}_2, \dots = \bar{X}_1, \dots \quad (12)$$

In practice it is customary to choose one of the methods of measurement as a standard method for which we would have the potentially infinite sequence, let us say

$$S_1, S_2, \dots S_1, \dots S_n, S_{n+1}, \dots S_{n+1}, \dots \quad (13)$$

to set it off from the others. Obviously this sequence in order to serve as a basis for comparison should be a random one in the sense that it is representative of a state of statistical control. Requirement (12) then reduces to

$$\bar{X}_i = \bar{S}' \quad (i = 1, 2, \dots i \dots) \quad (14)$$

Let us pause for a moment to examine some of the proposed standard methods of measuring let us say length. In this case the character of the reference standard is either some arbitrarily chosen physical object such as the Imperial Standard Yard and the International Prototype meter or some natural phenomenon such as the velocity of light.

First let us consider the requirement of randomness or statistical constancy of the standard sequence (13). Here we are back to the problem of setting up a statistical control technique and the choice of operationally verifiable criteria for control in the sense considered in Chapter I. As we have seen, in order to attain this objective the statistician must provide the scientist with all mathematical distribution theory that is necessary in order to make possible the choice of the most efficient control criteria for the particular case in hand. The layman might, of course, expect that it would be quite simple to set up such a standard series. A glance, however, at the control chart record, Fig. 18, for the measurement of the velocity of light should be sufficient grounds for believing that there apparently does not exist to-day

any random comparison series of observations for the measurement of length based upon the concept of a natural constant.

Now let us see what the situation is for comparisons in terms of arbitrarily chosen physical standards. Some very interesting results have recently been given<sup>1</sup> by J. E. Sears, superintendent of the metrology department of the National Physical Laboratory. In addition to the Imperial Standard Yard there are in existence at least four Parliamentary copies. Table 6 shows the observed differences in millionths of an inch between the length of the Imperial Standard Yard I and the copies P.C. 2; P.C. 3; P.C. 5; and P.C. VI. Sears claims that the observations on P.C. 3 in 1876 and those on P.C. 3 and P.C. 5 in 1892 are suspect and hence he argues that according to the results shown in this table the lengths of the bars P.C. 2, P.C. 3 and P.C. 5 have remained in close agreement with that of the standard. However, he points out that not only the evidence given in Table 6 but also other evidence which he cites

Comparison	Difference in Millionths of an Inch							
	1852	1876	1886	1892	1902	1912	1922	1932
P.C. 2 - I	+21	+36	-	+ 6	-	-23	-19	-39
P.C. 3 - I	-33	+57	-	+55	-	-49	-61	-111
P.C. 5 - I	-55	-33	-	+70	-	-43	-23	-47
P.C. VI - I	-	-	-3	-	-192	-215	-217	-234

Table 6

indicates that P.C. VI contracted over this period in an exponential manner so as to approach the asymptotic value of  $-228 \times 10^{-6}$  inch which it would seem that the bar had now reached. Sears points out that the bar P.C. VI was made several years after the others and argues that perhaps the reason why the change in length is noted only in the case of P.C. VI is that the others had reached a stable state before the measurements in the table above were taken. Of course, another explanation might be that the earlier bars, including the Imperial Standard, might have been shrinking at the same rate.

For our present purpose, the only point I wish to make is that there is some evidence for believing that the arbitrarily chosen physical standard

-----  
1. Science Progress, Vol. XXXI, Oct. 1936, pp. 209-235.

of length cannot be assumed to give a random test series at least until several years have elapsed and the results of intervening tests have been studied.

What is still more important is that, from an operational viewpoint, the faith that we have in any such standard series comes about in effect by comparing several series which for theoretical reasons of a physical nature we believe should give the same results. Thus, measurements by the standard I, as well as by any one of the copies is capable of giving an infinite sequence. Hence corresponding to measurements by the five standard bars we would have five sequences of the form shown in (15).

$$X' \xrightarrow{0} \begin{cases} \{S_{11}, S_{12}, \dots S_{1i}, \dots S_{1n}, S_{1n+1}, \dots S_{1n+i} \dots \\ \{S_{21}, S_{22}, \dots S_{2i}, \dots S_{2n}, S_{2n+1}, \dots S_{2n+i} \dots \\ \{ \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \\ \{S_{i1}, S_{i2}, \dots S_{ii}, \dots S_{in}, S_{in+1}, \dots S_{in+i} \dots \\ \{ \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \end{cases} \quad (15)$$

The concept of being able to make duplicate copies of a standard is therefore in a sense a requirement that the copies be interchangeable in terms of the sequences (15) which characterize them in an operational way.

Now let us ask: what is the essential difference between the operational representation of  $X'$  in (11) and that in (15)? The answer is that, from a theoretical viewpoint, any one of the methods of measuring shown in (11) might be chosen as a standard. If it were, there would be the question of producing duplicate standards and we would have a set of sequences of the type (15) for each method or, in other words, for each sequence in (11).

From the viewpoint of statistical theory, however, the requirements on the sequences in (11) are different from the requirements on the sequences in (15). For those in (11) and (15) we have the requirement (12) but for those in (15) we have the much broader requirement

$$f(S) dS = f_1(S_1) dS_1 = f_2(S_2) dS_2 = \dots = f_1(S_1) dS_1 = \dots \quad (16)$$

which is supposed to symbolize the requirement that the sequences in (15) may all be considered random samples from the same universe.

But if we are going to make the most efficient use of material, we must close up on the tolerances as far as it is economical to go. In this process, we must make use of two kinds of statistical criteria: a) those involved in the operation of control and b) those required to test the consistency of the sequences used in giving operationally definite meaning to the true value  $X'$  schematically illustrated in (11) and (15). Criteria under (b) are obviously those for testing significant differences in averages and for testing whether it is reasonable to believe that a given set of sequences came from the same state of statistical control. In other words, progress toward the ultimate goal of efficient use of raw materials by reducing tolerances to an economic minimum will necessarily involve extensive use and requirements of tests for significant differences.

Emphasis, however, should be placed upon the fact that in the use of statistical tests for significance it is necessary to use large enough samples to reduce to a satisfactory level the risks of making errors in judgments. The reason for such action is similar to that for going to a sample of 100 to 1000 in trying to establish a tolerance range, even in the simplest case of drawing from a normal bowl, as was pointed out in the discussion of Fig. 15. Also I think it is significant from the viewpoint of applying statistical theory, to note how extensive the series of measurements apparently have to be before we can hope to gain much from trying to analyze a set of data as though it were a sample from a bowl. For example, in the beginning of any investigation involving the measurement of any "true" value there are usually only a few methods of measuring known. At least in the field of physical and chemical science, the requirement of consistency between the results obtained by different methods has been a powerful influence in directing attention to the so-called constant errors. It would appear that, in general, it is of little value to make very large numbers of measurements by any one method until it has been found to give results that are more or less consistent with those of other methods. If, however, a large number of measurements are to be made as, for example, in the measurement of the velocity of light,

it would seem that much would be gained by applying statistical criteria of control for detecting assignable causes of variability for in no other way can we apparently reach the state of statistical control and maximum validity in prediction.

Someone may ask why go further than scientists have gone in trying to attain random sequences of physical measurements satisfying the criteria (12) and (16)? The answer would seem to be that, in just the same way as industrial applications of scientific principles have brought more and more stringent requirements as to accuracy in measurement, so will any further steps toward attaining maximum efficiency in the use of materials bring with them requirements on the method of measurement as to state of statistical control and maximum consistency both of which will necessitate the extensive use of statistical theory and technique.

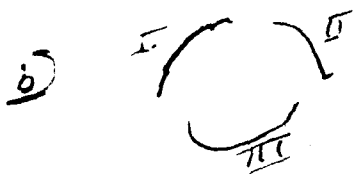
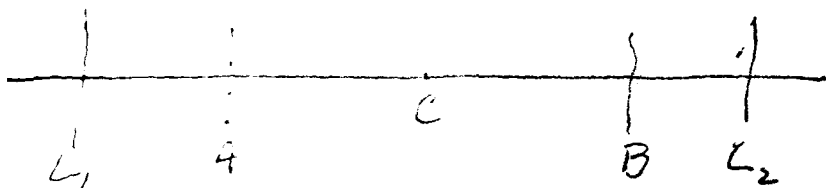
From what has been said in this chapter, it would appear that we must have a much more intimate knowledge of the properties of materials than we now have if the engineer of the future is to be able to minimize tolerances and thereby attain maximum efficiency in the use of materials. Furthermore it must be apparent that this ideal can only be attained by the application of statistical theory in establishing criteria for control and other criteria for testing consistency of data. When a state of statistical control is reached the method of setting up the tolerances becomes a purely statistical problem.

I.

# PRESENTATION OF RESULTS OF MEASUREMENT OF PHYSICAL PROPERTIES & CONSTANTS.

## † Review Comments

a) View point - Quality Control.  
MASS PRODUCTION



how can we tabulate data in  
this step to make the system.

giving statement on USE of Math. Statistics  
is in this process.

## II

Two USES of Data in form  
 $X = \Delta X$



v

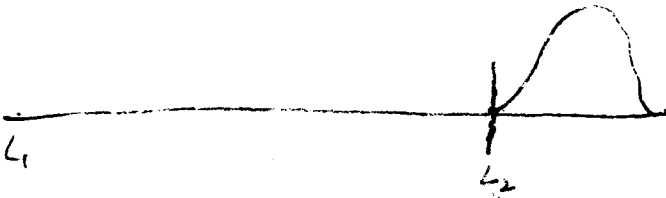
a) X as "best" value

$$1 \text{ metre} = 39.370147 \text{ inches}$$

Similar in a way to

$$\pi = 3.1416$$

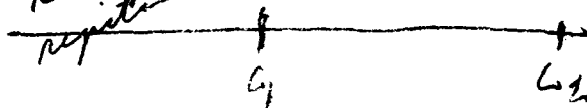
b) Contrast with problem



Not only set  $L_2$

But put in action limits to  
give quality Assurance.

c) or keep process  
repetitive min.



## III THREE CRITERIA SIGNIFICANCE

$\Delta x$  - can you do anything  
control action limits

$\Delta x$  - Can you sense it if  
you try

$\Delta x$  - Do you care  
like the question is to whether or  
no we should use relativity theory

2.  $2.99776 \pm 0.00004$  - Aug 1934  
 $2.99785 \pm 0.00005$  - Aug 1936  
 $2.9978 \pm 0.0002$  - from 1937

b)

$$L \times 10^{27} = 6.551 \pm 0.013$$

$$L \times 10^{27} = 6.547 \pm 0.008$$

Mod. of rupture

Prob

Species A  $5360 \pm 2252$

" B  $5365 \pm 780$

Which chosen

Significance -  $n_1 = 2000$   
 $n_2 = 4$

SS

1 metre = 39.370147 inches.

## CHAPTER III

### PRESENTATION OF THE RESULTS OF MEASUREMENT OF PHYSICAL PROPERTIES AND CONSTANTS

#### A Worthy Goal:

When you can measure what you are speaking about and express it in numbers, you know something about it, but when you cannot measure it, when you cannot express it in numbers, your knowledge is of a meagre and unsatisfactory kind.

Lord Kelvin

#### But:

..... knowing begins and ends in experience; but it does not end in the experience in which it begins.<sup>1</sup>

C. I. Lewis  
Harvard University

In the previous chapters, we have seen that the three steps in attaining economic control of quality are dependent one upon another. For example, we have seen that the problem of establishing tolerance limits that will make possible the most efficient use of materials can only be solved in the light of information obtained in the second and third steps of production and inspection or judgment of quality. In addition, there must be taken into account all pertinent data available in scientific and engineering literature and particularly those accumulated by standardizing bodies such, for example, as the American Society for Testing Materials. Literally thousands upon thousands of measurements are available in many instances while at other times we may have only a few. In both cases, however, the attempt is made to set up a tolerance range,

$$X = L_1 \text{ to } X = L_2$$

for any quality X such that in the course of production the probability p of producing a piece of product with the quality outside these limits satisfies the condition

$$p \leq p^*$$

---

1. "Experience and Meaning", The Phil. Review, Vol. XLIII, page 134, March 1934

where  $p'$  is some arbitrarily chosen value. Broadly speaking, this involves the problem of trying to summarize all pertinent information in the form of a range  $I \pm \Delta X$ .

This is perhaps the simplest type of problem of presentation of data, viz., the presentation of a series of  $n$  measurements  $X_1, X_2, \dots, X_n$ , of some physical constant or of some quality characteristic of a material such as the density of pure iron, tensile strength of steel, or the like. It is, of course, very common practice to summarize and report the results of such measurements in the form

$$X \tag{17}$$

or  $X \pm \Delta X$ . (18)

For example, I find in the Smithsonian Tables<sup>1</sup> the density of pure iron at ordinary temperature given as

$$7.86 \text{ gm/cm}^3.$$

Likewise, I find in another recent and authoritative table<sup>2</sup> the density at approximately the same temperature given as

$$(7.871 \pm 0.002) \text{ gm/cm}^3.$$

The problem of presentation of experimental data may be considered from one or the other of two viewpoints; a) from that of one who has some experimental results of his own or of others to present as scientific "facts" independent of how or for what purpose such facts may be used, and b) from that of one who wants to use such results previously presented. For example, on the one hand the worker in the field of "pure" science wants to present his own findings as objective facts. Likewise scientists often attempt to bring together the experimental facts obtained by others in the form of tables.<sup>3</sup> On

- 
1. Smithsonian Physical Tables, 8th revised Edition, The Smithsonian Institution, 1933, p. 160.
  2. Physical Constants of Pure Metals, The National Physical Laboratory, London, 1936, p. 6.
  3. For example, The Smithsonian Tables of Physical Constants; the International Critical Tables; Engineering Handbooks; Committee Reports of such societies as the American Society for Testing Materials, the American Chemical Society, and the like.

the other hand, professional men and engineers from many fields are typical of those who want to use the results thus presented.

We may, for example, pursue a subject such as physics or chemistry for what they are in themselves. In this sense we may study the structure and properties of materials; measure the fundamental constants of nature and discover the "laws" of science. As an artisan, however, we approach such subjects in an entirely different way - we have for example bio-physics, bio-chemistry, agricultural physics and chemistry, engineering physics and chemistry, and so on. Nevertheless, the "scientific facts" are presumably the same whether viewed by the pure scientist or the artisan. For example, the velocity of light  $c$ , the gravitational constant  $G$ , or any one of the seven so-called fundamental constants of nature is presumably independent of whether the observer is a pure or an applied scientist. In fact, tables of physical and chemical properties are apparently supposed to present the known facts for all.

Conceptually, of course, the density of pure iron in the units chosen is some single value which could be expressed in the form (17). If we could discover this true value with certainty, we could put it down once and for all as a fact. In practice, however, we cannot knowingly discover this true value - we can simply make measurements and draw inferences from the measurements thus made. Use of data involves the making of valid predictions or inferences about the future. From the viewpoint of such inferences, the original data constitute evidence and not facts. For example, there are three important aspects of knowledge from the viewpoint of presentation of data for use. These are schematically illustrated in Fig. 20.

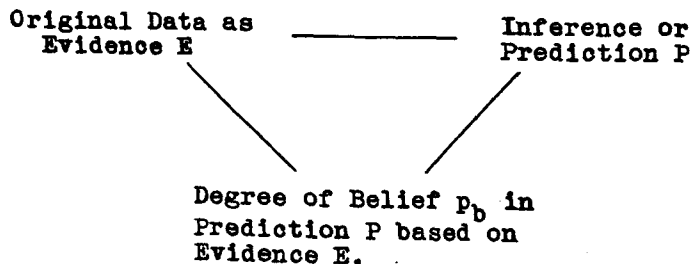


Fig. 20

To illustrate the practical significance of considering the problem of presentation of data from these three aspects let us consider the density of

pure iron as given above in the form  $(7.871 \pm 0.002) \text{ gm/cm}^3$ . Does the range as given provide us simply with a summary of the data or is it intended as an inference or prediction based upon the data? As a summary of the data, it might constitute evidence for several different inferences but, if it is intended as an inference, it may or may not be valid. Now in establishing tolerance ranges and particularly when one tries to reduce the range to an economic minimum, it is necessary to be able to establish inferences in the form of tolerance ranges that will be found valid. Much the same situation maintains in trying to render valid judgments about quality. It is therefore essential that we consider carefully the nature of the evidence that must be available in order to be able to render valid predictions or inferences in the form of tolerance ranges, for example. Obviously, however, there is nothing in a summary in the form of  $X \pm \Delta X$  that indicates whether or not tolerance limits or other inferences derived therefrom may be expected to be valid.

In what follows I shall try to show that under no conditions is a summary in the form of  $X \pm \Delta X$  alone an adequate basis for drawing valid inferences as to tolerance ranges or for rendering valid judgments as to the range of variability to be expected in quality fluctuations. Under certain conditions which are inherently statistical in nature, evidence in the form of  $X \pm \Delta X$ , together with the sample size, is adequate provided we have other evidence that these conditions have been fulfilled. In general, however, we shall see that such a summary is far from adequate - in addition we must have evidence as to the state of control of the quality and evidence as to whether or not the observed data are consistent with other pertinent data.

For example, let us turn to a table of physical constants such as the Smithsonian Tables.<sup>1</sup> On pages 103-107, we find approximately 130 ranges tabulated in the form  $X \pm \Delta X$ . Are inferences of a given kind drawn from these ranges of equal validity? Under what conditions and for what form of inference would it be possible to set down such a set of ranges that would have equal validity?

Ever since the time of Gauss, error theory has been used extensively

-----

1. loc. cit.

in summarizing data in the form of a range  $X \pm \Delta X$ . Several attempts have been made to apply some of the more recently developed statistical techniques.<sup>1</sup> Such applications have, however, been recently questioned by Norman Campbell<sup>2</sup>, von Friesen<sup>3</sup>, and others. It is hoped that the following discussion will throw some light on the question as to the role statistical theory may be expected to play in establishing such ranges.

Briefly then our problem is to consider ways and means of summarizing a series of observations in such a way as to make possible valid inferences in the form of tolerance ranges and quality judgments of a certain character.

#### FACING SOME PRELIMINARY DIFFICULTIES

Let us look at the tabulated densities of pure iron in the forms (17) and (18) above. What do they mean? One cannot expect to use such information intelligently unless one can answer this question. Is the "true" density  $7.86 \text{ gm/cm}^3$ ? Or does the true density lie somewhere within the range  $(7.871 \pm 0.002) \text{ gm/cm}^3$ ? No one knows the answer to such questions. Some one may suggest that the probability is  $p$  that the true value lies within this range. But what does this mean? Obviously, if one could discover the true value, he would find that it simply either did or did not fall within this range. As Bridgman has pointed out,<sup>4</sup> there does not seem to be any method of verifying the probability of a single event such as we have here.

When data are tabulated in the form  $X \pm \Delta X$  there seems to be pretty general agreement in calling  $X$  an estimate of some true value. Now what about  $\Delta X$ ? Let us look at some of the things it is called in recent literature:

1. Probable error<sup>5</sup>
2. Estimate of probable error<sup>5</sup>

- 
1. Deming, W.E. and Birge, R.T. "On the Statistical Theory of Errors," Rev. of Mod. Phys., Vol. 6, No. 3, July, 1934.
  2. Campbell, Norman, "The Statistical Theory of Errors," Proc. of the Phys. Soc., Vol. 47, pp. 800-809, 1935.
  3. Sten von Friesen, "On the Values of Fundamental Atomic Constants," Proc. Roy. Soc. Lon., Series A. No. 902, Vol. 160, pp. 424-440, June 1937.
  4. Bridgman, P.W., "Statistical Mechanics," Bull. Amer. Math. Soc., Vol. XXVIII, #4, 1932, pp. 225-245. The Nature of Physical Theory. Princeton University Press, 1936.
  5. Many current articles and tables.



3. "Even-bet" error.<sup>1</sup>
4. Estimate of reasonable limit of error.<sup>2</sup>
5. Degree of uncertainty.<sup>3</sup>

Is the practical significance of  $\Delta X$  the same irrespective of what it is called?

For example, suppose we look up the recorded velocity of light in vacuo. Table 7 gives values in units of  $10^{10}$  cm/sec suggested by three authorities after making surveys of the literature as of the dates given.

<u>Birge<sup>4</sup> (1934)</u>	<u>Bond<sup>5</sup> (1936)</u>	<u>von Friesen<sup>6</sup> (1937)</u>
2.99776 ± 0.00004	2.9978 <sub>5</sub> ± 0.0000 <sub>5</sub>	2.9978 ± 0.0002

Table 7

The ranges here given are respectively termed from left to right "even-bet", probable error, and reasonable limits of error range. In what sense do these three kinds of range differ? What useful purpose, if any, does one range serve a practical man that the other two ranges do not serve?

Next let us pass on to a consideration of another difficulty; viz., that different analysts using the same set of data do not always arrive at the same values of  $X \pm \Delta X$  even though they call them the same. Schematically the situation may be represented as follows:

$$\begin{array}{l} \text{Original Data in} \\ \text{the form of } n \text{ observed} \\ \text{Values.} \end{array} \xrightarrow{H_1} (X \pm \Delta X)_1$$

where  $\xrightarrow{H_1}$  represents the operation of going from the given data to the tabulated range  $(X \pm \Delta X)_1$  for the person  $H_1$ . Since this operation of deriving a given kind of range, as, for example, the probable error range, is not the same

1. "On the Values of Fundamental Atomic Constants", by R.T. Birge, The Physical Review, Vol. 52, No. 3, 241, August 1, 1937.

2. S. von Friesen, "On the Values of Fundamental Atomic Constants", Proc. Roy. Soc. Lon., Series A, Vol. 160, pp. 424-440, June 1937.

3. International Critical Tables, First Edition, 1926, page 17.

4. "The Velocity of Light", Nature, Vol. 134, page 771, Nov. 17, 1934.

5. "Fundamental Physical Constants", Phil. Mag., Series 7, Vol. xxii, pp. 624-632, October 1936.

6. "On the Values of Fundamental Atomic Constants", Proc. Roy. Soc. Lon., Series A, No. 902, Vol. 160, pp. 424-440, June 1937.

The subscript 5 in Bond's result is supposed to indicate variability in the fifth decimal place.

for all operators or analysts and since, in general, the range depends upon the operation used, it is pretty difficult to see how a tabulated range can be used as a basis of valid prediction without first knowing the operation. For example, the value of  $X$  appearing in the range  $(X \pm \Delta X)_i$  is very often the arithmetic mean of the  $n$  observed values; it is just as likely, however, to be a weighted mean of some kind where the weighting and the choice of mean depends upon the analyst. Likewise the  $\Delta X$  is often derived in many different ways as, for example, from the mean deviation, standard deviation, range and certain other statistics of a sample. Then too there is no universally used technique for getting a  $\Delta X$  from a given statistic. Finally, it should be noted that special grouping of the data is often followed.

For example, Fig. 21 shows the recorded ranges for several different determinations of the velocity of light as given in a recent article.<sup>1</sup> The length of the vertical line in each case is proportional to the recorded range. One might be tempted to compare these ranges visually and conclude that

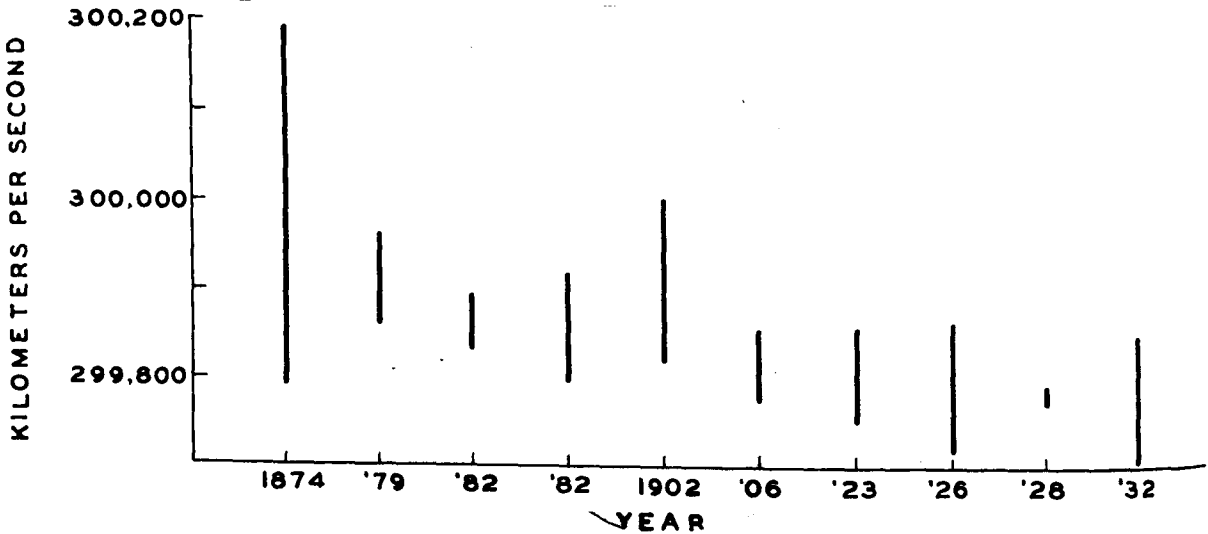


Fig. 21

scientists have been doing some pretty good work in closing up on the range within the last few years. But when one looks at the original data and finds that the operation of getting the ranges have not been the same in each case - in fact that they have been very far from the same - how can he justify any comparisons until he has first gone back to the original articles?

1. "Values of Fundamental Atomic Constants", by S. von Friesen, Proc. Royal Soc. London, Series A, No. 902, Vol. 160, pp. 424-440, June, 1937.

Perhaps it is not out of place here to give an example of how such tabulated ranges are sometimes compared in the literature. In a recent paper by Eddington<sup>1</sup>, he puts the question: Suppose I have occasion to use Planck's constant and I find in reference books two determinations,

$$h \times 10^{27} = 6.551 \pm .013$$

$$h \times 10^{27} = 6.547 \pm .008.$$

Assuming that these are to be taken at their face value, which one shall I choose? He argues that the latter is the more useful to him because it limits h to a narrower range and hence will lead to sharper conclusions.

Before commenting on this example let us consider another which in many ways at least is very similar and let us see if the same line of argument would likely be used. Suppose that an engineer is interested in building a pole line and he has before him the following modulus of rupture figures on two kinds of poles, expressed in the same units:

Species A: 5340 ± 2252

Species B: 5365 ± 1803

Assuming that these data are to be taken at their face value, which species should be chosen? Should the engineer choose the latter on the assumption that poles of this species will be found on an average to have about the same tensile strength as the others and to show a smaller dispersion? Suppose now that the engineer finds out that although the limits were set in the same way for both species, there were over 2000 poles for species A and only five for species B. Would the choice now remain the same?

Going back to the illustration given by Eddington, I am not sure what the expression "Assuming that these are to be taken at their face value" covers. Assuming, however, that the choice is valid upon the basis of this conditioning phrase, I do not recall having seen any practical example where I would feel free to make the choice suggested. In other words, I do not know of any instance where such a qualifying assumption would be justified. This

-----  
1. Eddington, A.S. "Notes on the Method of Least Squares" Proc. of the Phy. Soc., Vol. 45, Part 2, #247, pp. 271-282, 1933.

is certainly true for the ranges shown in Fig. 21 and similar ranges tabulated for the fundamental physical constants.

These are but typical of the difficulties that are continually coming to my attention in the course of my every day work and also in the course of my work as Chairman of the Joint Committee on the Development of Statistical Applications in Manufacturing and Engineering sponsored jointly by certain scientific and engineering societies.<sup>1</sup> Scientists and artisans alike have long looked to the theory of errors and least squares to throw light on the problem of presentation. With the developing emphasis during the last few decades upon statistical theory and its applications in one field or another, it has been quite natural that these developments would also be expected to shed light on the problem. I hope to show as we proceed some of the important ways in which statistical theory may be used to advantage in considering the problem of presentation. It will, however, be necessary to call attention to some fundamental and, so far as I know, unsolved difficulties, the solution of which seems to fall outside of anything we may hope to obtain from the application of formal statistical theory, at least as it is ordinarily considered.

SOME FUNDAMENTAL CONSIDERATIONS

Let us assume that we have a set of N normal "bowl-universes" for which we know the expected values  $\bar{X}'_1, \bar{X}'_2, \dots, \bar{X}'_1, \dots, \bar{X}'_N$ . Given samples  $n_1, n_2, \dots, n_1, \dots, n_N$  from the respective bowls, we have seen in the previous chapter how "Students'" theory enables us to calculate a range for each bowl such that  $p_1^N$  of the ranges thus calculated may be expected to include the corresponding values  $\bar{X}'_1, \bar{X}'_2, \dots, \bar{X}'_1, \dots, \bar{X}'_N$ . Such a range we shall call a fiducial range and such a prediction we shall symbolize  $P_1$ . Similarly we have also seen how we may, by taking a large enough sample from each bowl, establish for each bowl a valid tolerance range. The establishment of such a range<sup>2</sup> we shall refer to as a prediction,  $P_2$ . In what follows we shall be primarily

- 
1. The American Society of Mechanical Engineers, the American Society for Testing Materials, the American Statistical Association, the American Mathematical Society, and the Institute of Mathematical Statistics.
  2. These two kinds of predictions are illustrated in Figs. 14 and 15 respectively of Chapter II.

concerned with the tolerance prediction  $P_2$  but shall contrast it with the other type  $P_1$  which has been considered so extensively in recent literature.

It is obvious that the measurements denoted by the X's may be measurements of the quality X for a set of objects or a series of measurements of a physical constant.

Next we should note the sense in which we shall consider that a series of measurements of some quality characteristic differs from a series of drawings at random from an experimental bowl. The series of n numbers in the two cases representing a series of n observations may be identical and representable in such a case by

$$X_1, X_2, \dots X_1, \dots X_n \tag{19}$$

In the case of the bowl, the whole of our information is contained in the set of n numbers as numbers. In the case of the n physical measurements, however, the situation is fundamentally different as schematically illustrated below,



where for each X there is an associated physical condition, all of which may or may not be "essentially the same."

Now as we pointed out in a previous chapter, it is very common practice in summarizing data, particularly the measurements of research such as those on the velocity of light - to assume that the conditions are essentially the same, which assumption we may symbolize by the expression

$$C_i = C_j, \tag{21}$$

it being kept in mind that the C's do not represent numbers but "conditions" in the sense we use that term in the phrase, "the same essential conditions."

From the viewpoint of prediction in the physical case, both the X's and the C's must be taken into account, unless the C's satisfy condition (21) in which case they may be neglected and the practical case reduces to the bowl case and thus becomes a purely statistical problem. We have seen, however, that even in the case of refined physical measurements, the set of X's when

ordered in time do not, in general, satisfy the control chart criterion. From the viewpoint of presentation, therefore, it appears that we are seldom justified in handling a set of data as though they constituted a sample drawn from a bowl unless we have evidence that they also satisfy certain criteria of statistical control. Hence if an engineer or scientist be given predictions of either type  $P_1$  or  $P_2$  without knowing whether or not the data satisfy the criteria of control, he is in no position to determine how much belief he should place in the predictions which, of course, would be valid if the conditions are such as maintain in the case of the bowl. Since one is not interested in predictions that he doesn't know whether or not to believe, it would seem to follow that in tabulating and summarizing data it is necessary to provide evidence as to whether or not the observed values satisfy some criterion of control in the sense of the previous chapter.

Furthermore, one is seldom interested in the measurements as such but instead is interested in what they tell about some so-called objective true value. Operationally this means that the prediction in terms of one method of measurement must be consistent with predictions in terms of other known methods of measuring. Hence in trying to determine the degree of belief to place in predictions of types  $P_1$  and  $P_2$ , in so far as they reveal objectivity, one needs to have evidence as to the degree of consistency.

It follows from such general considerations that the presentation of physical data (20) should not be treated the same as the presentation of data drawn from a bowl even when the practical data arose under a state of statistical control and are consistent with other pertinent data. Evidence for believing that the physical data arose from such a state and are consistent with other pertinent data is necessary before one reading the presented results can determine how much belief he should place in the predictions.

Now we are in a position to take up the matter of summarizing data in each of several cases, starting with the simplest and working up to the most difficult. To begin with, it is desirable to distinguish between two fundamental conditions: a) when the  $n$  observed values are known to arise under a state of statistical control and b) when we do not know that the data arose under such conditions.

DATA FROM STATE OF STATISTICAL CONTROL

Perhaps our nearest approach to such a state of control is that of drawing with replacement from an experimental bowl-universe. We shall at least assume that an experimental sample from a bowl satisfies the requirement that the order in which the numbers are drawn is of no significance from the viewpoint of prediction, or, in other words, that the whole of the information or evidence given by a sample of n from such a bowl is contained in the frequency distribution of the numbers.

To begin, we should differentiate between the problem of summarizing the set of numbers as numbers and that of summarizing them for the purposes of making a specific kind of prediction. It is often desirable in order to save space, to try to summarize a frequency distribution of n finite numbers in terms of a set of m numbers,  $\Theta_1, \Theta_2, \dots, \Theta_m$ , where  $m < n$  and the  $\Theta$ 's are determined from the sample. The ideal aimed at is to secure a set of numbers,  $\Theta$ 's, such that one could go from the  $\Theta$ 's to the X's as well as going from the X's to the  $\Theta$ 's. This ideal we may represent schematically as follows:

$$X_1, X_2, \dots, X_1, \dots, X_n \rightleftharpoons \Theta_1, \Theta_2, \dots, \Theta_m. \tag{22}$$

Now, of course, unless  $m=n$  it is not feasible to attain this ideal. If, however, we tabulate the arithmetic average  $\bar{X}$  and the root mean square deviation  $\sigma$  of the n numbers along with the sample size n, we can always look backward from the summarized results and say with certainty<sup>1</sup> that not more than

$$\frac{n}{t^2}$$

values of X are outside the limits

$$\bar{X} \pm t\sigma .$$

For example, let us consider the frequency distribution of the sixteen numbers given in Table 3 of the previous chapter: 6.661; 6.661; 6.664; 6.667; 6.667; 6.671; 6.672; 6.672; 6.674; 6.675; 6.676; 6.678; 6.679; 6.681; 6.683. Given only the average  $\bar{X} = 6.672$  and standard deviation  $\sigma = .0061$  of this set

-----  
1. Cf. Tchebycheff's Theorem, page 95, Economic Control of Quality of Manufactured Product, by W. A. Shewhart.

of 16 numbers one could say with certainty without ever having seen the original distribution that not more than  $\frac{16}{t^2}$  of these numbers lie outside the limits  $\bar{X} \pm t\sigma = 6.672 \pm t \times .0061$ . Tchebycheff's theorem applies as a description of the distribution observed in the sample, irrespective of how the numbers are distributed so long as they are all finite. In this sense it is a remarkable theorem, but, of course, it does not allow one to differentiate between distributions having the same  $\bar{X}$ ,  $\sigma$ , and  $n$ . Hence if the use of the data summarized in the form of  $\Theta$ 's involves inferences which depend upon the distribution of the numbers in the sample it is always necessary to go beyond the three statistics  $n$ ,  $\bar{X}$  and  $\sigma$  in the process of summarization.

Thus far we have considered the problem of summarizing the data from the viewpoint of being able to go from the summary in terms of the  $\Theta$ 's back to the original distribution. This phase of the subject is seldom if ever considered by statisticians, because the assumption is customarily made that we simply wish to go from the sample to the "population" or "statistical universe" which in the case of a sample drawn from a bowl would simply be the true distribution of numbers in the bowl. Schematically the situation is as illustrated in Fig. 22. In other words, if the distribution in the bowl can be approx-

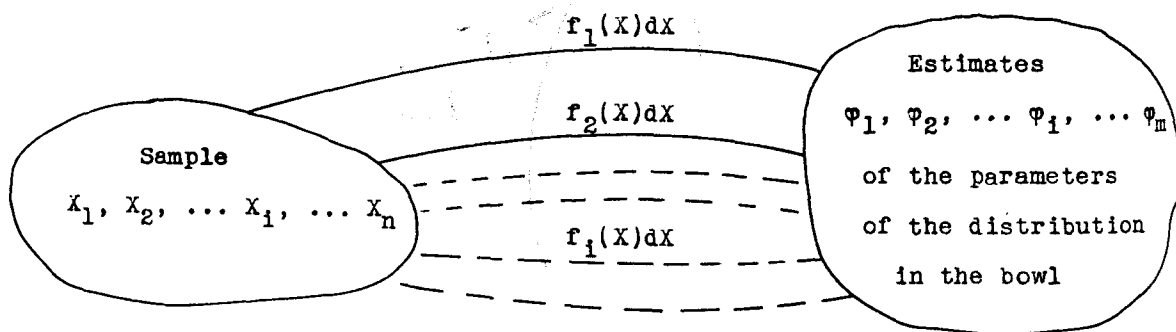


Fig. 22

imately represented by a continuous frequency function  $y = f(x)$  involving  $m$  parameters, the statistician usually conceives of his problem as estimating these parameters.

From the viewpoint of presentation, the point I wish to make is that the statistically "best" estimate depends upon the assumed form of the function  $f$ . For example, let us consider estimates of the standard deviation  $\sigma'$  in the



bowl. These can usually be expressed as a function of the standard deviation  $\sigma$  of the sample. Thus if  $f(x)$  is normal, the estimate customarily accepted as the best is  $\sqrt{\frac{n}{n-1}} \sigma$ . But for some other forms of  $f(x)$  the estimates expressed as a function of  $\sigma$  would be different. Thus in order to make estimates we must first either know or assume as known the form  $f(x)$  of the distribution in the bowl. For different functional forms these may, as it were, be different paths from the sample to the estimates. Much the same situation holds for predictions in terms of fiducial and of tolerance ranges. Hence, in general, if a sample from a bowl is summarized in terms of estimates,  $\varphi_1, \varphi_2, \dots, \varphi_1, \dots, \varphi_m$ , of parameters or in terms of predictions without stating the assumption made as to  $f(x)$ , there is no way of using the information contained in the summarized form as a basis for making other estimates upon the basis of other assumptions which for one reason or another it may be desirable to make at some later time. Except in the case which almost never occurs in practice where the function  $f(x)$  is known, the justification for any assumed  $f(x)$  must rest either upon the evidence provided by the sample or upon that and prior experience. Hence the assumption as to  $f(x)$  is always subject to change with additional information, but if the past information given by a sample is available only in the form of estimates, we cannot make use of this information in making other estimates. Hence in commercial work at least it is often desirable to summarize data in terms of a set of  $\Theta$ 's that will give the most information possible about the distribution in the sample and that do not involve any assumptions about the distribution in the bowl. In general, it is possible to choose the  $\Theta$ 's in the form of symmetric functions that quite satisfactorily meet these requirements. Then one can use tabulated data in quality reports and the like to test various assumptions. We may summarize what has just been said schematically as follows:

$$X_1, X_2, \dots, X_1, \dots, X_n \longrightarrow \Theta_1, \Theta_2, \dots, \Theta_m \begin{array}{l} \xrightarrow{f_1(x)} \varphi_1, \varphi_2, \dots, \varphi_s \\ \xrightarrow{f_1(x)} \text{Predictions } P_1, P_2 \end{array} \quad (23)$$

where the  $\longrightarrow$  is supposed to picture one method of going from the sample to  $\Theta$ 's and the  $\xrightarrow{f_1(x)}$  many ways of going from  $\Theta$ 's to either the  $\varphi$ 's or the  $P$ 's depending upon the choice made by the analyst.

Now let us proceed to a consideration of some special cases.

Case I. Distribution in Bowl Normal

To make the problem concrete let us consider the following sample of four:

1.7; 0.2; 1.4; 0.5

The best way to summarize such a sample in terms of  $\Theta$ 's is to take

$$\Theta_1 = \bar{X} = .950 \quad \text{and} \quad \Theta_2 = \sigma = .619$$

Since in this case the distribution in the bowl is normal, it is completely determined in terms of the average  $\bar{X}$  and standard deviation  $\sigma$  of the distribution. Now the best estimates of these are usually considered to be:

$$\varphi_1 = \bar{X} = .950 \quad \text{and} \quad \varphi_2 = \sqrt{\frac{n}{n-1}} \sigma = .715$$

"Students'" theory enables one to calculate the fiducial range corresponding to any previously chosen value of probability  $p$  in the form:

$$\bar{X} \pm t\sigma$$

where  $t$  can be determined from tables.

A tolerance range may also be set up in terms of  $\bar{X}$  and  $\sigma$ , as we saw in the previous chapter, but one cannot judge from the tabulated tolerance range alone how valid it is. To arrive at an estimate of its validity, we must know at least,

$$\bar{X}, \sigma, \text{ and } n.$$

Hence even in the simplest case we should tabulate the sample size if the data are to be used at any time in establishing tolerances. The striking thing is, however, that in this simple case of drawing from a normal bowl, we do not need to tabulate more than  $\bar{X}$ ,  $\sigma$  and  $n$  irrespective of the size of the sample.

Case II. Form of Distribution in the Bowl Known but Not Normal

Let us start with a consideration of what can be done with the summary of the information contained in the sample in the form of the average  $\bar{X}$ , standard deviation  $\sigma$  and sample size  $n$  of the sample. Suppose we use these statistics to establish fiducial ranges by the same mathematical rule that we used in the case of the samples from a normal distribution. Should we expect

appreciably the same degree of validity in ranges thus set up as we found in the case of the normal law?

Fig. 23 shows 100 fiducial ranges for a probability of .5 for three forms of experimental universe: normal, rectangular and right triangular. The ranges in order to be valid should include the corresponding true values in the bowls on an average of 50 times in 100, with a deviation from this ratio no greater than can be attributed to sampling fluctuations.

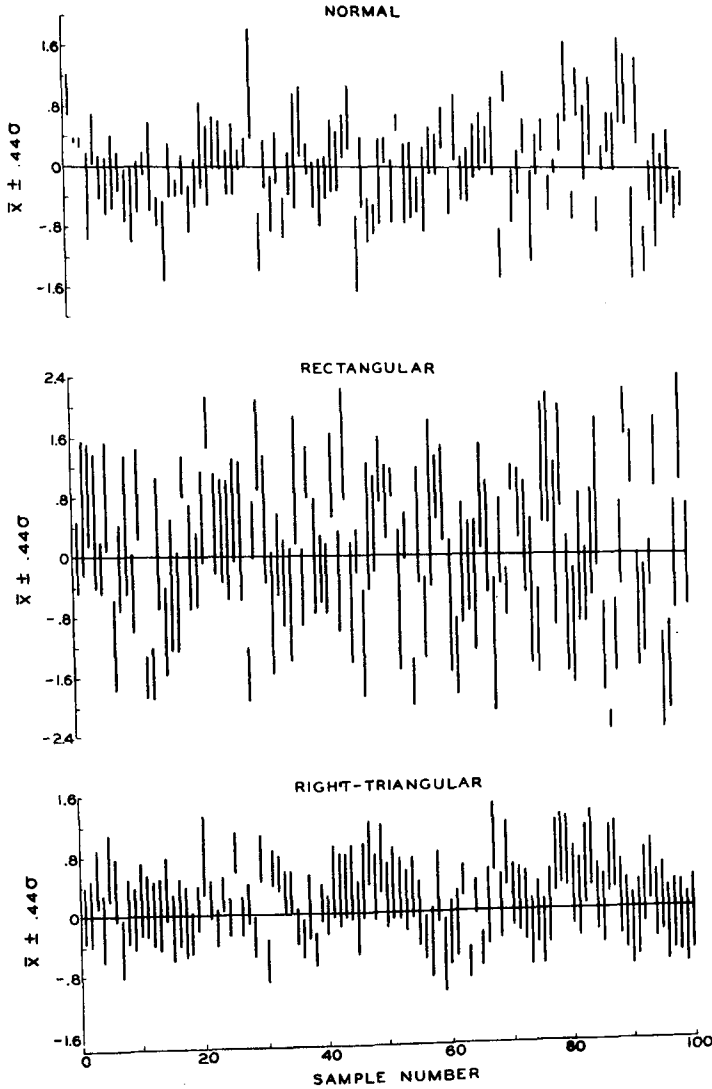


Fig. 23

The record of inclusions as seen from the figure are:

Normal

51

Rectangular

56

Right Triangular

68

These results illustrate how the degree of belief in a prediction of type  $P_1$  must depend upon the state of our knowledge of the shape of the universe even when they differ no more from normality than the two here chosen.

It would be difficult to stress the practical significance of this point too much. If every time one had a sample of 4 from a bowl, he were to make a prediction of type  $P_1$  corresponding to a probability of .5, in just the way he would if he knew the universe to be normal, he might at least expect to be off something like the difference between .50 and .68. In the case of so-called "multimodal" universes the errors might even be much larger.

Under such conditions, I do not know of any better way to summarize a sample of  $n$  observed values from a bowl than in terms of  $\bar{X}$ ,  $\sigma$ , and  $n$  if we are limited to three numbers in the summary. But the point that should be stressed is that in presenting data in such summary form for the use of others one certainly would not be justified in simply giving  $\bar{X}$ ,  $\sigma$ , and  $n$  if he also knew the functional form  $f$  of the distribution in the bowl, because the one who may later use the data may thereby be led into serious pitfalls. In such a case one should give:

$\bar{X}$ ,  $\sigma$ ,  $n$  and  $f$ .

One should note, however, how inefficient such a summary may be from the viewpoint of setting tolerance limits for values of  $p$  of the order of magnitude of .003 or less. Suppose, for example, that we have given:

$\bar{X} = -.0028$ ;  $\sigma = .9663$ ;  $n = 1000$ ;  $f$  is right triangular.

I do not know of any way of setting up such a tolerance range in terms of such a summary that would approach in validity, one that might be set up on the basis of the observed distribution in the sample of 1000 shown in Fig. 24. For example, I think almost everyone will agree that tolerance limits -1.4 and 2.6 would satisfy the requirement. It is interesting to contrast with this example of setting tolerances, that of setting the same tolerance limits upon the basis of a sample of 1000 from a normal bowl. In Fig. 15 of Chapter II, we saw how successfully this could be done in terms of simply  $\bar{X}$  and  $\sigma$ . As was pointed out in Chapter II, it is necessary to have a large sample - perhaps even 1000 - before one feels justified in trying to set minimum tolerance limits even when it is known that the sample is from a normal bowl. For cer-

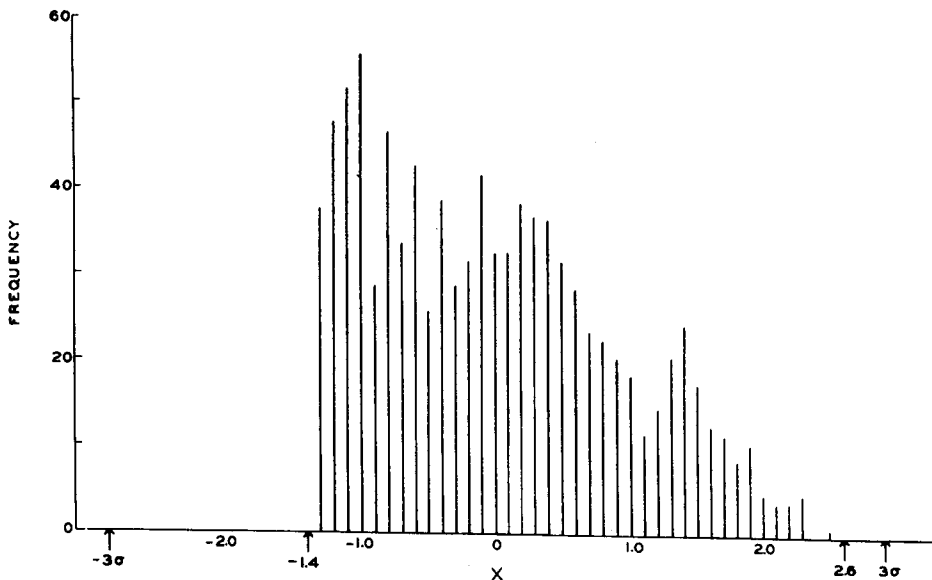


Fig. 24

tain functional forms of the universe, even larger samples would be required. In the present state of our knowledge of the theory of estimation and the establishment of valid ranges of variability - particularly tolerance ranges - in terms of a few  $\Theta$ 's, I feel that one is not justified in trying to summarize samples of the size required as a basis for establishing valid tolerances in this way unless he knows that the universe is normal.

Case III. Form of Distribution in Bowl Unknown

Whereas in the previous case it is a pretty difficult problem to summarize the information given by a sample in terms of  $\Theta$ 's unless the universe from which the sample came is normal, it is obviously much more difficult to do so when the form of the universe is unknown. As pointed out above, we must know the form of  $f$  in order to make a valid prediction. But if we do not know  $f$ , we must assume a form upon the evidence given by the sample.

To illustrate, let us consider the following sample of eight drawn from a bowl

- 1.7; 10.7; .2; 1.4; 10.0; 10.4; 0.5; 10.6

How would you summarize these numbers? Would you be satisfied with  $\bar{X}$ ,  $\sigma$  and  $n$  as a basis for determining how much belief  $p_b$  you would have in a prediction of type  $P_1$  of the type that would be valid for a sample from a normal universe? Suppose we plot these eight values on a straight line, Fig. 25.

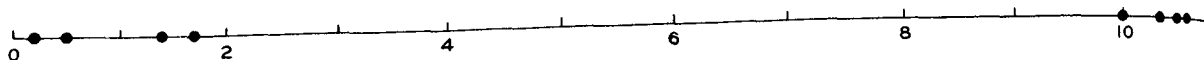


Fig. 25

I assume that no one would have as much faith in such a prediction of type  $P_1$  where all he knows about the universe in the bowl is that given by the sample shown in Fig. 25 as he would have if he knew the universe were normal.

Anyone familiar with even elementary sampling theory appreciates the fact that a comparatively large sample - something like a sample of 1000 or more - must be available even when drawn from a bowl before one can place much reliance in his judgment as to the functional form  $f$  of the distribution in the bowl, particularly if one is interested primarily in the tails of the distribution. Furthermore, such a person is familiar with the serious difficulties of trying to judge the form of the distribution when the only information available as to the distribution is in a set of statistics,  $\Theta$ 's. Hence, from the viewpoint of summarizing a sample drawn from a bowl in which the form of the distribution is unknown, it does not - at least in engineering work - and particularly the setting of tolerances - appear desirable to give a summary in the form of  $\Theta$ 's. This is particularly true if the sample is large. If we symbolize by  $f_0$  the observed distribution in the sample then it appears that it is desirable to present

$$f_0, \bar{X}, \text{ and } \sigma .$$

#### Some General Comments from the Viewpoint of Practice

In practice the statistical state of control represented by drawings from a bowl is the limit approached in the process of removing assignable causes of variability. It is therefore the condition for which tolerances can be set in such a way as to make possible the most efficient use of materials. Hence it is significant to note in the light of discussion of this section the following points:

1. Even after we have attained a state of statistical control a comparatively large sample is required in order to provide a basis for making valid predictions.
2. A summary of a set of  $n$  data in the form of a range  $\bar{X} \pm \Delta X$  is not in itself sufficient grounds for establishing a tolerance range. In all cases it is necessary to record the sample size  $n$ .
3. When the functional form  $f$  of the universe is known, we should give at least  $\bar{X}$ ,  $\sigma$ ,  $n$  and  $f$ , and sometimes it is desirable to include the observed distribution  $f_0$  in the sample. When the functional  $f$  of the universe is unknown, we should give at least  $f_0$ ,  $\bar{X}$  and  $\sigma$ .

Some comment should perhaps be made at this point as to what may seem to be a conflict between these three suggestions and the current practice of tabulating on the one hand and the tendency of modern statistics toward the use of small samples and the emphasis on the advantages of summarizing data in terms of efficient statistics. The conflict is more apparent than real. The differences arise primarily because of the varied uses to be made of the tabulated results. In practice we are not justified in assuming a statistically controlled state, in general, and in finding assignable causes of variability small samples can be used to advantage. For the purposes for which such samples are used, the information can usually be summarized satisfactorily in terms of  $\bar{X}$ ,  $\sigma$  and  $n$ . However, in order to minimize tolerance limits it is desirable that we know the distribution function for each quality characteristic in terms of either:

$$f_0(x) \text{ or } f(x) dx,$$

where  $f_0(x)$  stands for the observed distribution in a sample of at least 1000 and  $f(x)$  symbolizes a satisfactory approximation in the form of a continuous function.

#### PRESENTATION OF DATA - CUSTOMARY CONDITIONS

In the customary case, a sample of  $n$  observed values of some quality or some physical constant do not give evidence of having arisen from a state of statistical control. Observed values obtained by one method of measurement are usually found to be assignably different from similar values obtained by other methods of measurement. Such a state of affairs complicates the problem of

trying to summarize data far and beyond the complications found in the different cases of summarizing a sample of data from a bowl. Furthermore, the problem of presentation as considered in this section cannot be solved by the statistician acting alone as was the case in the previous section.

Case IV - Need for Evidence as to State of Control

Let us assume that we are given a sequence of n measurements,  $X_1, X_2, \dots, X_1, \dots, X_n$ , of some quantity X. We have already noted that to every  $X_1$  there is a condition  $C_1$  in the sense that we use the term condition in the phrase "same essential conditions". Every  $X_1$  has its handle  $C_1$

$$X_1 - C_1$$

and both must be taken into account from the viewpoint of summarizing data. Broadly speaking, the scientist or engineer must take account of the  $C_1$  and the statistician must be responsible for handling the X's at least in the limiting case where the X's behave as though they came from a state of statistical control. Now, of course, there is in general no quantitative way of expressing  $C_1$ . All we can perhaps hope to do is to depend upon the scientist to suggest hypotheses about the conditions - in particular that certain of them are essentially the same. For example, I presume that Heyl's measurements of G in the appropriate units might be set down in the form shown in Table 8. From the viewpoint of a statistician, this would constitute an hypothesis to be tested.

Gold	Platinum	Glass
6.683 - $C_1$	6.661 - $C_2$	6.678 - $C_3$
6.681 - $C_1$	6.661 - $C_2^2$	6.671 - $C_3^3$
6.676 - $C_1$	6.667 - $C_2^2$	6.675 - $C_3^3$
6.678 - $C_1$	6.667 - $C_2^2$	6.672 - $C_3^3$
6.679 - $C_1$	6.664 - $C_2^2$	6.674 - $C_3^3$
6.672 - $C_1$		

Table 8

If, however, the scientist chose to present such a set of data as though the  $C$ 's were all the same, the contributions of the statistician would be limited by such a form of presentation. Of course, even under these conditions, the statistician can test this hypothesis if the data are given in the sequence in which they were taken. It may be of interest to know that taking these 16 measurements in the order in which they were presumably observed and applying Criterion I for control, we get no evidence of lack of statistical control,

*less like  
"intentional  
error"*



whereas when the data are given in the form of Table 8 one is almost certain that they did not arise under a condition of statistical control. Here we have a good illustration of the importance of the C from the viewpoint of tabulation.

In general, one seldom if ever finds that the first n measurements of a physical quality or constant (or the quality X of the first n pieces of product turned out by a given process) satisfy Criterion I for control. The state of statistical control is only approached through the process of discovering and weeding out assignable causes. This fact coupled with the fact that experience indicates that the set of assignable causes can be found and removed is significant from the viewpoint of presenting data in that evidence of assignable causes found and removed from time to time adds to the belief that one has in the end results representing a state of statistical control.

For example, Fig. 26 shows a control chart for 16 averages of 4 measurements on a running sequence of 64 measurements of resistance on as many pieces of a new kind of product. This evidence is consistent with the assump-

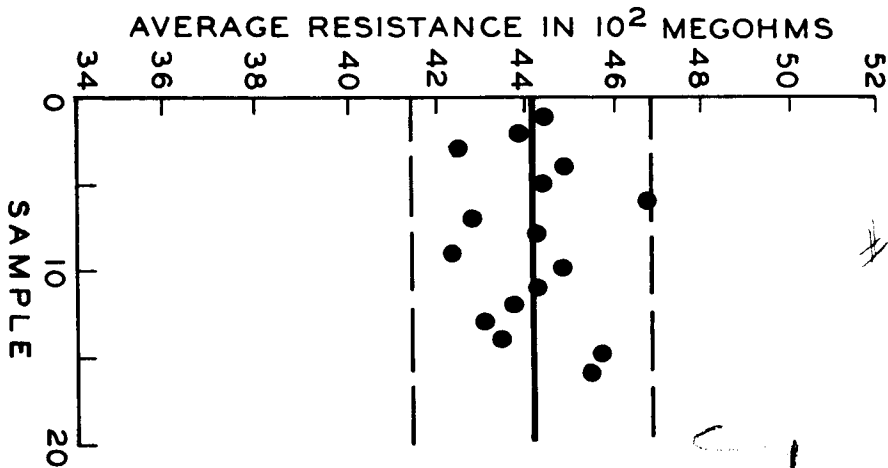


Fig. 26

tion that a state of statistical control has been reached. Now, let us look at a similar chart shown in Fig. 27 which gives the averages for the first 51 samples of four. Given this additional information together with the statement that certain assignable causes were found and removed from the process between the time the data in Fig. 27 and those in Fig. 26 were taken, I think one's belief that the data of Fig. 26 represent a statistically controlled state is strengthened. This fact is of particular importance from the viewpoint of

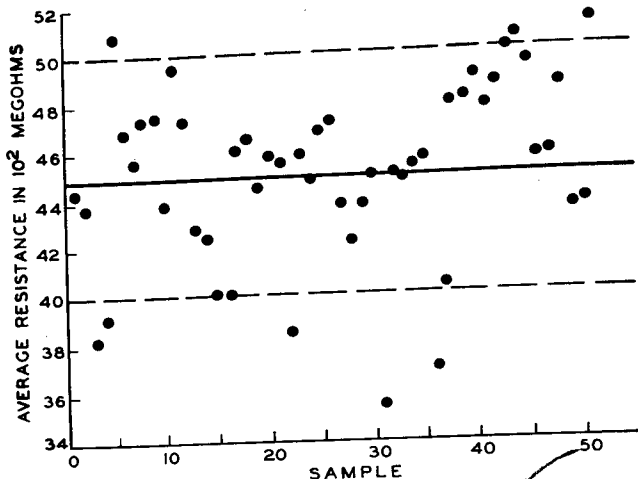


Fig. 27

keeping a running report on quality as a basis for judging quality, in so far as it shows how such a report may indicate progress toward the attainment of a state of control even though such a state has not been attained.

In the process of testing data for evidence of control, I have shown elsewhere why it is desirable for the scientist or engineer to divide the original data into comparatively small groups which he thinks arose under the same essential conditions. These are then tested for control by some criterion involving in general the use of the average  $\bar{X}$ , standard deviation  $\sigma$ , and sample size  $n$  of each subgroup. Suppose, however, that one wished to continue the study of the resistance of the new kind of material considered above until he had sufficient evidence for setting valid, minimum tolerance limits. Beginning at about the data shown in Fig. 26 and continuing on until a sample of something like one thousand is reached, the data should be kept in the form of a frequency distribution, for reasons set forth in the previous section. Here, in other words, we see the difference between summarizing data for getting evidence of control and summarizing data apparently coming from a state of statistical control for the purpose of providing a basis for establishing tolerance limits that will make possible the most efficient use of material.

There is, however, another problem which we should consider, namely that of setting tolerance limits when no attempts at statistical control are being made. In this case, and for more or less obvious reasons, the maximum and minimum observed values play a very significant role in enabling an engineer

to set tolerance limits that will take in most of the product, although such limits do not permit of making the most efficient (but not necessarily the most economic) use of material. Particularly is this true if a large number of measurements representing a wide range of conditions are available. For example, the 20,000 measurements of the tensile strength of malleable iron from 17 different sources shown above in Table 4 of Chapter II constitute a good example. There are reasons which we need not go into here for presenting the average also so that we may say that under such conditions the following statistics should be tabulated

Max., Min.,  $\bar{X}$ , and  $n$ ,

if we are limited to four.

#### Case V. Need for Evidence as to Consistency

Let us consider as an example the setting of tolerance limits on the measurement of a physical constant such as the velocity of light. As previously pointed out in Chapter II, the problem is the same analytically as that of setting tolerance limits on the true value of quality of pieces of product of a given kind. It is true, of course, that the tolerance limits on a quality must take into account not only the variability of the "true" quality but also that of the method of measurement. Hence the problem of setting tolerances on the measurement of a presumably constant value of a given quality always constitutes a part of the job of setting tolerances on a quality characteristic.

Suppose that one is given in the appropriate units the average  $\bar{X}$ , standard deviation  $\sigma$ , and sample size  $n$  for the measurements of the velocity of light previously considered:

$$\bar{X} = 299,773.85; \quad \sigma = 13.37; \quad n = 2885$$

Let us also assume, although contrary to fact, that these data satisfy Criterion I of control and that the distribution is approximately normal. Would we be justified in using this set of data alone as a basis for setting tolerance limits for the measurement of the velocity of light? Obviously the answer to this question is "No", if by measurement we are to include measurement not only by the method used in this case but also by other methods admitted by scientists as having a just claim for consideration.

For example, let us compare this set of measurements with another set

of 651 more recently reported by Anderson. Fig. 28 shows<sup>1</sup> control charts placed end to end for the two series and constructed as best one can<sup>2</sup> from the data as recorded. The striking thing to note is that the two averages are sig-

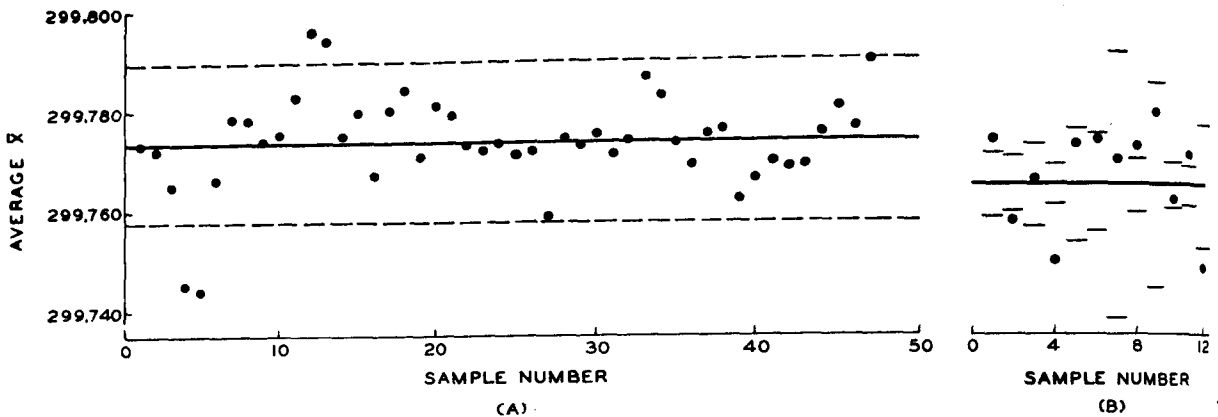


Fig. 28 - Measurement of Velocity of Light - A) By Michelson, B) By Anderson  
nificantly different. For example, Anderson's data give:

$$\bar{X} = 299,764.15; \sigma = 14.96 \text{ and } n = 651.$$

The ratio of the observed difference in averages to the estimated standard deviation of this difference is

$$\frac{\bar{X}_1 - \bar{X}_2}{.6370} = 15.23.$$

It is indeed very unlikely that a difference so large as this would arise as a result of random sampling. Incidentally, I think it is this general type of experience in which it is found that different test methods appear to give assignably different results that leads scientists to stress the importance of checking for consistency between measurement by different methods rather than to stress repetition of the same measurement a great many times.

Obviously the kind of evidence that one would want to have before trying to set a minimum tolerance range in a given case would be the maximum observation given by the method producing maximum values in general and the

1. "Measurement of the Velocity of Light in a Partial Vacuum", by Michelson, Pease, and Pearson, Astrophysical Journal, 82, 1935 (2885.5 observations); "A Measurement of the Velocity of Light" by W. C. Anderson, Physical Review, 8, July, 1937 (651 observations).
2. Anderson records average deviation for each sample; the sigmas used in the control chart equal the mean deviations multiplied by  $\sqrt{\pi/2}$ .

minimum value for the method producing minimum values. One would also want to know perhaps the number of different methods of measurement that had been tried because "constant errors" have in the past usually been discovered through the use of different methods of measurement. If one takes the time to look back through the literature in physics, let us say, for a period of some twenty years or more, he will find quite a variation in the accepted values for many of the constants there tabulated. The same is true for measurements of the atomic weights in chemistry as is illustrated in Table 9 which shows the accepted values relative to oxygen = 16.00 for the dates 1931 and 1936.

International Atomic Weights

<u>Element</u>	Relative Atomic Weight	
	Oxygen = 16	
	<u>1931</u>	<u>1936</u>
Arsenic	74.93	74.91
Cesium	132.81	132.91
Columbium	93.3	92.91
Iodine	126.932	126.92
Krypton	82.9	83.7
Lanthanum	138.90	138.92
Osmium	190.8	191.5
Potassium	39.10	39.096
Radium	225.97	226.05
Ytterbium	173.5	173.04

Col. 2 From Table 595, Smithsonian Physical Tables, 8th ed., Ed. by Fowle, 1933, Washington, D.C.

Col. 3 Published by the Journal of the American Chemical Society, vol. 58<sup>1</sup>, 1936.

Table 9

From the viewpoint of establishing a tolerance upon such measurements, it therefore appears that the following information provided by the available data are of major importance:

Maximum, Minimum, and K

where K is the number of different methods involved. Certainly not very much information is provided by a weighted average and an estimate of a so-called probable error so long as the results given by different methods are assignably different. Perhaps in this case more than in any other, the names of the scientists are also a necessary factor. It would seem therefore that statistical theory does not contribute much to the technique of presenting evidence upon which to base a tolerance range under conditions that are not statistically

*is from ? statistician  
in the group as a whole*

controlled. However, if for some reason or other it becomes necessary to close up on such a tolerance range by detecting and eliminating all constant errors, statistical tests for significant differences would become a necessary tool in the process.

NOTE ON THE TABULATION OF PHYSICAL PROPERTIES

Thus far in the discussion, emphasis has been laid upon the problem of tabulating data as a basis for establishing a tolerance range. Let us look now at the customary practice of tabulating a range  $\bar{X} \pm \Delta X$ , as in the case of the Smithsonian Tables. It is my understanding that certain scientists maintain that approximately 50% of the tabulated ranges should include the corresponding true values. As already pointed out earlier in this chapter, I do not see how such a true value is operationally verifiable. All that we can ever hope to do is to take further measurements by one or more methods. But so soon as we think of the problem from this viewpoint it is the concept of tolerance range and not that of fiducial range that becomes of importance.

Suppose, however, that one could discover the "true" values in some way. At least such an assumption is operationally justifiable in the theoretical sense. A prediction  $P_1$  that 50% of the true values would be found to lie within their respective ranges is valid only if the data used as a basis for setting the ranges constitutes a random sample from a normal universe. In other words, the conditions under which the samples were taken should be in a state of statistical control and the observed values should be free from constant errors, neither of which condition is likely to be found in practice. Under such conditions what justification would we have in expecting such a prediction to be valid? If it is subject to error, what information do the data give as to the magnitude of this error? Unless one can answer these questions satisfactorily, he will not likely place much confidence in the validity of such ranges.

Again let us consider another common practice of tabulating the probable error of the average and not of a single observation. The dominating idea is that by making the sample size larger and larger we are getting closer and closer to the true value  $\bar{X}$ ' in the same way as we do when taking the average of a sample from a bowl. For example, if we take the standard deviations

PENALTY FOR PRIVATE USE TO AVOID  
PAYMENT OF POSTAGE, \$300.

UNITED STATES DEPARTMENT OF AGRICULTURE  
BUREAU OF CHEMISTRY AND SOILS  
WASHINGTON, D. C.  
OFFICIAL BUSINESS

22

By messenger

Dr. W. A. Shewhart

Hotel Washington

628

W

set in such cases solely  
by statistical rule.

II <sup>Q</sup> Is it possible to design  
<sup>an</sup> experiment that the  
applicability of statistics  
can be safely assumed  
without direct proof?

Ans Proper design is  
necessary to obtain data  
that can be characterized by a  
model tool. But I do not  
see how design can itself  
can assure that the same  
essential conditions will  
be maintained in the future.



Q: Statistical methods should not be used until it has been proved that they apply. Is the statement justified?

Ans. A statistician calculates one or more statistics from a set of data -  $O_1, O_2, \dots$ . Then if he <sup>assumes that he</sup> knows the functional form of the ~~same~~ population from which he can derive the distribution

$$f(\theta) d\theta \text{ in samples}$$

Then he makes prediction of let us say Type 1 or 2.

This is valid if he knows that future observed  $O$ 's

I have for years been interested  
in the use of many of the  
techniques now considered in  
the theory of design. Five years  
of my experience was on the  
study of the physico-chemical  
properties of the granular carbon  
used in the microphone. Design  
was an important factor ~~but~~  
getting to the place where  
I could proceed <sup>with</sup> finding and eliminating  
assignable causes. The  
elimination however came through  
the process of repetitions.

III What direct test is

Will you please emphasize this point, implied in your Tuesday lecture, on the use of small samples:

A small sample may validly be used to test the existence of a given characteristic in the population sampled, say the fraction defective, but it may not be validly used to determine the fraction defective in the population.

Thus we may, under certain conditions calculate the chance of occurrence of a particular fraction defective in a small sample from a population with specified fraction defective. We may not, in general, however, estimate the fraction defective in the population from that of the small sample; certainly we may not attach any probability statement to such an estimate.

I may throw six dice once and validly test the existence of a  $1/6$  probability of the occurrence of a "one", but I may not argue that solely because there were three "one's" in my set of six dice that the chance of the occurrence of a "one" is  $1/2$ , or any particular value.

of the averages for the two sets of 2885 and 651 observations of the velocity of light shown in Fig. 28, we have

$$(.0046 \text{ and } .0230) \text{ km. sec}^{-1}$$

respectively, and this in the face of the fact that the observed difference in averages is  $9.70 \text{ km. sec}^{-1}$ . Furthermore as we have seen, both sets of data give evidence of lack of control. Under such conditions what use could be made of a range in the form  $\bar{X} \pm \Delta X$  where you know that  $\Delta X$  is an estimate of probable error but you do not know the sample size  $n$ , for this is usually omitted from tabulations? In any case it would be much more useful to tabulate

$$\bar{X}, \sigma \text{ and } n$$

and then one could estimate the probable error not only for this sample size but for any other.

Now, of course, in the customary case the estimates of probable error are derived by combining measurements by different methods and by special groupings. Certainly if one were going to check predictions made in this way he would have to make further measurements and the predictions would hold, in general, only in the case that each of the samples of data used in the check were made up from data obtained by the different methods used in getting the original estimate and in the same proportion. Such a form of sample is, however, exceedingly artificial in character.

#### SUMMARY COMMENT - SIGNIFICANCE IN QUALITY CONTROL

Now let us try to gather together some of the pertinent conclusions and indicate the importance of these from the viewpoint of keeping a running quality report that will form an adequate basis for giving quality assurance and for minimizing tolerances.

We started this chapter with two quotations: In the words of Lord Kelvin, "When you can measure what you are speaking about and express it in numbers, you know something about it." So it is in science and engineering we try to measure the physical qualities of the things with which we work and express the results in numbers  $X \pm \Delta X$  as though that constituted knowledge. But back of that pair of numbers there is a set of data or measurements. That set of data constitutes a bit of experience. The process of knowing begins in this

experience but it does not end there. This experience is but a bit of history. The process of knowing ends in experience but an experience in the future - another set or other sets of data yet to be taken if you please. The process of knowing does not end in the numbers  $X \pm \Delta X$ . From the viewpoint of knowing, the numbers are either a part of a prediction or a part of the evidence on which a prediction is made. We shall now review the four ways in which numbers obtained in measurement enter into a report.

As Original Data

Three ways are available for presenting a series of  $n$  measurements  $X_1, X_2, \dots, X_n$  of a physical quality or constant:

1.  $X_1, X_2, \dots, X_1, \dots, X_n$  and H.  
 $\begin{array}{cccc} | & | & | & | \\ C_1 & C_2 & C_1 & C_n \end{array}$
2. Sequence ordered in time.
3. Observed frequency distribution,  $f_0(x)$ ,

where the C's stand for conditions in the sense of the phrase "the same essential conditions", and H stands for the person judging the conditions. The C's cannot be expressed in numbers. The person H can only provide hypotheses about the C's which may be tested in terms of the X's. If the scientist judge H judges the conditions to be the same, the preceding discussion shows that the data should be tabulated in the form of a sequence, permitting of a check on the assumption in terms of the  $n$  values of X, by means of Criterion I or some other criterion of control. If such data are tabulated only as a frequency distribution, they then become statistics<sup>1</sup> and not physical measurements. Only if we knew that the data arose under a state of statistical control would we be justified in presenting original data in this way. This represents the limiting state beyond which human judgment or insight as to the causes of variation cannot go. Since, however, as we have seen, most data show evidence of lack of control, they should always be presented by the experimentalist H in either one or the other of the first two forms.

-----  
 1. In the lay sense in which this term is used.

### As $\Theta$ 's

The  $\Theta$ 's as we have seen stand for functions of the original data that are independent of any hypothesis as to the functional distribution  $f$  of the "universe" of which the data may be considered a sample. The  $\Theta$ 's summarize some of the characteristics of the original frequency function  $f_0(x)$ , and therefore are limited in the same way as  $f_0(x)$  is limited. The sample size  $n$  should always be given. Sometimes the average  $\bar{X}$  and the standard deviation  $\sigma$  of the distribution  $f_0(x)$  are sufficient. Sometimes instead of  $\sigma$  we should tabulate the maximum and minimum observed values and sometimes  $f_0(x)$  should also be given.

### As $\varphi$ 's and $P$ 's

The  $\varphi$ 's as we recall are estimates of universe parameters and although derived from the original data, involve an assumption as to the functional distribution of the universe from which the original  $n$  values of  $X$  are assumed to come. They differ from the  $\Theta$ 's in the fact that although for one set of  $X$ 's there can only be one set of  $\Theta$ 's, there can be many sets of  $\varphi$ 's depending upon the assumptions made. The  $\varphi$ 's in this sense constitute interpretations or predictions of a certain kind based upon the original data. The  $\Theta$ 's do not involve interpretations. They constitute a partial summary of the numerical facts completely presented by  $f_0(x)$ .

The predictions  $P_1$  and  $P_2$  involving fiducial ranges and tolerance ranges are still more involved forms of interpretation. Reasons are presented for believing that the conditions of <sup>maximum</sup> validity for  $P_1$  are seldom satisfied.

### As Evidence

*Anyone*  
*Even a fool*  
*may predict*  
ones. Hence a practical man is always like the proverbial man from Missouri in respect to a prediction until he sees the evidence. Three factors have been stressed in addition to the human element represented by the scientist  $H$  who is responsible for the hypotheses about the physical conditions represented by the  $C$ 's. These are

1. Quantity of information as represented by sample size  $n$ .
2. Evidence for control including a history of the assignable causes found and removed.

### 3. Evidence for consistency.

What role does the statistician play in the problem of presentation? The answer is that he is a co-partner with the scientist or engineer as long as the latter judges the C's to be different. When the scientist or engineer is ready to give up the ghost and says that  $C_i = C_j$  - the conditions are essentially the same, the statistician steps in. He begins as a doubting Thomas and tests for statistical control which he seldom finds. If, however, he does find the conditions apparently controlled as they would be in a statistical state, then he and he alone is in a position to make the requisite valid predictions.

Finally, let us ask: What has all this to do with quality control? In the first chapter, we got a picture of the inter-relatedness of the three fundamental steps in control. There and also in the second chapter we saw the need for a record of quality measurements not only from the viewpoint of giving quality assurance but also from the viewpoint of providing in the end an adequate basis for establishing tolerance limits that will make possible the most efficient use of materials. Such information must be made available in a running report. In so far as the principles considered in this chapter are applied in the production of such a report, one has gone as far as possible in making full use of the available data. In the preparation of such a report, the engineer and statistician must play cooperative roles.

## CHAPTER IV

### SPECIFICATION OF ACCURACY AND PRECISION

The concept is synonymous with the corresponding set of operations.<sup>1</sup>

P. W. Bridgman  
Harvard University

The successful operation of cotton mills is likewise becoming a business of precision.

ROBERT B. WEST, President  
Riverside and Dan River Cotton Mills

### INTRODUCTION

We are told that necessity is the mother of invention. It is true that when man became a measuring animal he had to adopt standards of length, mass, and the like. Commerce and industry called for the legalizing of certain standards and the establishment of methods of measuring with requisite accuracy and precision in terms of such standards. Likewise, the introduction of interchangeability about 1787 necessitated accurate measurement and the invention of gauges. The steady increase in the accuracy of interchangeable parts produced under manufacturing conditions has led to the invention of standard length gauges with .00001 inch tolerances and pushed the accuracy of test methods out to .000001 inch.<sup>2</sup> Both pure and applied science have pushed farther and farther the requirements of accuracy and precision.

Applied science, particularly in the mass production of interchangeable parts, is becoming perhaps even more exacting than pure science in certain matters of accuracy and precision. It undertakes to make large numbers of things of a given kind with specified degrees of these factors, such as accuracy of 1% or precision of 1%. Failure to meet the requirements may mean rejection and accompanying increase in the cost of production. Such specifications may become the basis of contractual agreement, and any indefiniteness in

-----

1. The Logic of Modern Physics, The Macmillan Co., N.Y., 1928, p.5.

2. Cf. Gauges and Fine Measurements, by F. H. Rolt, The Macmillan Company, London, 1929, Vol. 1, p.10.



the meaning of the terms accuracy and precision used therein or methods of measuring the same may lead to misunderstandings and even legal action. The development of modern methods of mass production to specification is the mother of many changes in our concepts and use of the terms accuracy and precision. The object of the present paper is to set forth some of these changes necessitated by economic production practices.

For example, let us consider a specification of the form:

- A) The accuracy of the test method shall be  $\pm 1\%$ .
- B) The precision of the test method shall be  $\pm 1\%$ .

Under such conditions when is one justified in saying:

- a) The accuracy of this test method is  $\pm 1\%$ .
- b) The precision of this test method is  $\pm 1\%$ .

Now suppose that a consumer makes the specification A and B and a producer makes the claim a) and b). How would you as an independent and unbiased observer or scientific judge, as it were, proceed to verify the producer's statement a) and b)?

I suppose a layman has the right to assume that if anyone ever attempts to say just what he means and mean just what he says, perhaps that one should be a scientist or engineer when specifying accuracy and precision and when making statements such as a) and b) involving these terms. What we shall have to say, therefore, is of interest not only as a consideration of the special problem of specifying accuracy and precision but as a consideration of the limit to which one may hope to go in saying just what one means in a way that is subject to verification - something that is basic to all specification. For example, I think it will be readily admitted that the limit to which we can go in specifying the quality of a physical thing in a verifiable manner certainly depends, among other things, upon the limit to which we can go in specifying one simple quality characteristic, such as length, density, or the like, of that thing in terms of quantitative measurements of such quality characteristics.

At the beginning, therefore, it is perhaps well that we adopt some criterion by which we shall judge the meaning of the terms accuracy and precision.

OPERATIONAL CRITERION OF MEANING

This is not the place to discuss the present status of the operational theory of meaning. It would be difficult to say just where such a theory had its origin but it would perhaps be generally admitted that it represents the development of the theory and technique of saying what we mean and meaning what we say in a way that is subject to verification. Thus in the theory of errors and the theory of statistics we introduce such terms as true value, equally likely, population, population parameters, random, and the like. For example, in the theory of errors we generally postulate a true value  $X'$  of the thing being measured and define the error  $e$  of an observed value  $X$  by the equation

$$e = X' - X. \tag{24}$$

Now admittedly we do not know  $X'$  and perhaps can never know it. For example, if the measurement  $X$  is an observed value of the velocity of light and I make the statement that this particular value is in error by a certain amount, how would one proceed to verify this? The difficulty in doing this is apparent.

In the form in which the operational theory of meaning was first given broad consideration, the basic principle or criterion of meaningfulness seems to be the following: the meaning of a statement is the method of its verification. This is much the same criterion as Bridgman<sup>1</sup> adopts in his *Logic of Modern Physics*, 1928, and also is the form adopted by so-called logical positivists up to that time. In this sense, therefore, terms such as true value, equally likely, and the others listed above, would be considered as meaningless unless some one could devise means of verifying them. From the viewpoint of such a criterion, some argue that any statement about the probability of a single event such as the turning up of a head in the throw of a coin is meaningless. For example, let us consider the statement: The probability that the penny which I hold in my hand will turn up head when I throw it, is  $1/2$ . Let us contrast this statement with the one: When I throw the penny which I hold in my hand, it will turn up heads. Now suppose that I toss the penny and it turns up heads. I have in a certain sense verified the second

-----

1. Loc. cit.

statement but in what sense have I verified the first one? It is difficult to say in so far as a single event is concerned, that the word probable has any verifiable meaning.

Under these conditions, there is today an increased effort on the part of many writing in the theory of probability and statistics to eliminate, in so far as possible, terms which are troublesome in the sense here under consideration. For example, some<sup>1</sup> recent books define probability without using the term equally likely. In this way they attempt to reduce the theorems of probability to formal mathematics. In general, however, just so soon as we eliminate such terms from discussions in the analysis of data, we tend to eliminate the terms which are intended to suggest at least the applicability of mathematics to practical problems. If, in the field of experience, we cut loose, as it were, from the use of all terms such as true value, random, and the like, which cannot be experimentally verified in a rigorous and absolute sense, we might seemingly hope to make statements which are subject to experimental verification. If, for example, we write specifications of accuracy and precision in such experimentally verifiable terms, it would presumably be possible to state without ambiguity what is meant in a way that could be verified. Such a possibility has, as already noted, a very important appeal to those charged with the writing of specifications. To be more specific, it might appear feasible to write a specification of the form A mentioned above, namely, the accuracy of the test method shall be 1%. Any statement such as (a) or (b) above involving the term is usually implies, I think, something about the characteristics of the test method not yet experienced. It involves, in other words, a prediction and is not simply a report of a past verification which can serve simply as history, as it were. In this sense such a statement may or may not turn out to be false. The practical significance of a statement of such a character obviously depends not simply on the form of the statement but upon the belief that we may place in the validity of the statement.

Another objective, of course, in adopting the operational theory of meaning is to eliminate from scientific discourse terms which have a purely

-----

1. See for example, Elements of Probability, by Levy and Roth, Oxford Univ. Press, 1936.

**Tafel 1. Allgemeine Konstanten**

$\pi$	3.14159 26535 89793 23846 26433 83279 50288 41971 69399 37510 58209 74944 59230 78164 06286 20899 86280 34825 34211 70679 82148 08651 32823 06647 09384 46095 50582 23172 53594 08128 48111 74502 84102 70193 85211 05559 64462 29489 54930 38196 44288 10975 66593 34461 28475 64823 37867 83165 27120 19091 45648 56692 34603 48610 45432 66482 13393 60726 02491 41273 72458 70066 06315 58817 48815 20920 96282 92540 91715 36436 78925 90360 01133 05305 48820 46652 13841 46951 94151 16094 33057 27036 57595 91953 09218 61173 81932 61179 31051 18548 07446 23799 62749 56735 18857 52724 89122 79381 83011 94912 98336 73362 44065 66430 86021 39501 60924 48077 23094 36285 53096 62027 55693 97986 95022 24749 96206 07497 03041 23668 86199 51100 89202 38377 02131 41694 11902 98858 25446 81639 79990 46597 00081 70029 63123 77381 34208 41307 91451 18398 05709 85
-------	--

$1:\pi$	0.31830 98861 83790 67153 77675 26745 02872 40689 19291 48091 29
$\sqrt{\pi}$	1.77245 38509 05516 02729 81674 83341 14518 27975 49456 12238 7
$1:\sqrt{\pi}$	0.56418 95835 47756 28694 80794 51560 8
$\log \pi$	0.49714 98726 94133 85435 12682 88290 89887 36516 78324 38044 24461 34053 6
$\ln \pi$	1.14472 98858 49400 17414 34273 51353 05871 16472 94812 915
$e^\pi$	23.14069 26327 79269 00572 90863 67948 54738 02661
$e^{-\pi}$	0.04321 39182 63772 24977 44177 37171 72801 12757 28109 81063
$e^{\frac{\pi}{4}}$	2.19328 00507 38015 45655 97696 59278 73822 34616 37642 0
$e^{-\frac{\pi}{4}}$	0.45593 81277 65996 23676 59212 94728 02941 94166 04365 238
$\sqrt{2}$	1.41421 35623 73095 04880 16887 24209 69807 85696 71875 37694 80731 76679 7380
$\sqrt{3}$	1.73205 08075 68877 29352 74463 41505 87236 69428 05253 81038 06280 55806 97945 2

Vielfache von $\pi$		Vielfache von $(1:\pi)$	
1	3.14159 26535 89793 23846 26433 83279 50	1	0.31830 98861 83790 67153 77675 26745 03
2	6.28318 53071 79586 47692 52867 66559 01	2	0.63661 97723 67581 34307 55350 53490 06
3	9.42477 79607 69379 71538 79301 49838 51	3	0.95492 96585 51372 01461 33025 80235 09
4	12.56637 06143 59172 95385 05735 33118 01	4	1.27323 95447 35162 68615 10701 06980 11
5	15.70796 32679 48966 19231 32169 16397 51	5	1.59154 94309 18953 35768 88376 33725 14
6	18.84955 59215 38759 43077 58602 99677 02	6	1.90985 93171 02744 02922 66051 60470 17
7	21.99114 85751 28552 66923 85036 82956 52	7	2.22816 92032 86534 70076 43726 87215 20
8	25.13274 12287 18345 90770 11470 66236 02	8	2.54647 90804 70325 37230 21402 13960 23
9	28.27433 38823 08139 14616 37904 49515 53	9	2.86478 89756 54116 04383 99077 40705 26

by Judge Quality in resp  
to interest

Inspection spec.

Facts

$x_1$	...	$x_n$	
$c_1$		$c_2$	H

III ACCURACY

SPECIFICATION  
of  
ACCURACY & PRECISION

**I JOINT COMMITTEE etc.**

- a) If you have question a Committee can be of service to you - write me.
- b) If you have question and if the ~~letter~~ <sup>letter</sup> are published I shall be glad to clarify my comments when published.
- c) References to specific examples ~~etc~~  
Drop Mr. Doby & write.

**II INTRODUCTION**

- a) Some aspects.  
This is the main point.
- b) Again look at three steps.

Specify  $L_1, L_2$   $p \leq p'$ .

" Element in  $n$  space

Specify means of getting adequate evidence.

c) Fundamentally precision and accuracy

d) Accuracy and precision shall be 1%

e) Accuracy and precision is 1%  
Step 1  
Step 2

### III THREE: PRECISION dist

a)  $x_1, x_2, \dots, x_n$  |  $\mu$

Well try  $\bar{x}$  and  $\sigma$  for  $n=4$   
.....

Random is more than we can define even in limited range.

General Terms ? or Spec

emotive meaning, such, for example, as the words "good" and "truth" and the phrase "hard cold facts". Thus we often see in the discussion of the analysis of data comments that such and such a formula applies only to good data. Also we run across such phrases as "the truth of the matter is that". I suppose that those who try to popularize science or who try to advertise some new product are most often accused of using terms primarily of emotive significance.

Students of the operational theory of meaning have within recent years, of course, begun to appreciate some of the serious pitfalls in trying to reduce scientific writing to a consideration of terms that can in practice be verified. They have, accordingly, broadened the meaning of verification to include so-called theoretical verification. Under these conditions we shall take the following as a criterion of meaning as a background for considering accuracy and precision:

Every sentence in order to have definite cognitive meaning must be theoretically verifiable or confirmable as either true or false upon the basis of evidence theoretically obtainable by carrying out a definite and previously specified operation in the future. The meaning of such a sentence is the method of its verification.

It is hoped that enough has been said to suggest at least the nature of the fundamental problem involved here in trying for the sake of definiteness to break away from the "objectivity" of the physical world except in so far as it can be expressed in terms of operationally verifiable characteristics. It would appear that such a path might lead to an authoritarian choice which would have some at least of the characteristics of authoritarian statements in general. A question which we must keep in mind, therefore, is: how far can one attempt to go in attaining the ideal of an operationally verifiable specification without including in the specification some person or group of persons and the experience which they must have had before making the specifications that are to be adopted.

#### CONCEPTS OF ACCURACY AND PRECISION

These two terms have been and continue to be used by all technical people in the discussion of both pure and applied science - they are among the most common in scientific literature. But try to find out just what either term means or the difference between them and, as is so often true with terms of common parlance, the meanings are not quite clear cut. In fact, the two



terms are given as synonyms in some of the best dictionaries.

Books on the theory of errors and the theory of measurements also use them often as synonyms. The author of one of the most widely known books on the precision of measurements, however, bemoans the fact that the two terms are used thus carelessly and indiscriminately.<sup>1</sup> By precision or precision measure of a result he refers to what he terms the best numerical measure of its reliability which can, as he says, be obtained after eliminating or correcting for all known sources of error. By the accuracy of a result he refers to the degree of concordance between it and the true value of the quantity measured. These definitions, however, introduce several terms and phrases such as "reliability" and "concordance" which are not adequately explained.

Obviously, if the terms accuracy and precision are synonymous, one cannot readily see any difference between the specification of an accuracy of 1% and a precision of 1%. If, on the other hand, we were to adopt the definitions made by Goodwin we would first have to define reliability and concordance and numerical measures of these before we could make intelligent specifications. It would appear, therefore, that one of the first things to be done is to decide whether or not the concepts are, as it were, synonymous. I personally feel that the concepts which these two terms try to define are fundamentally different in so far as they have applicability to the affairs of every-day life in engineering and science. We shall, however, a little later see that these terms have tended to become confused in the literature of the theory of errors and statistics because of certain assumptions that have been made to simplify the theoretical problem. Abundant evidence will be mentioned to show that in practice there is little justification for believing that the assumptions hold and hence there is little reason for confusing the two concepts from the viewpoint of practice.

Etymologically the term "accurate" has a Latin origin meaning "to take pains with" and refers to the care bestowed upon a human effort to make it what it ought to be. Likewise, "accuracy" in common dictionary parlance implies freedom from mistakes or exact conformity to truth. "Precise", on the other hand, has its origin in a term meaning "cut off", brief, concise. Like-

---

1. Goodwin, H. M., Precision of Measurements and Graphical Methods, McGraw-Hill Book Company, New York, 1913, pp.7-8.

wise, precision is supposed to imply the property of determinate limitations or that of being exactly or sharply defined.

In what follows, therefore, I shall take as a starting point the following distinction:

1. Accuracy in some way or other involves the concept of a difference between what is observed and what is true.
2. Precision involves the concept of reproducibility of what is observed.

Let us see what can be done to reduce this broad distinction to more definite terms.

Let us consider two very simple cases - one where the true value is unknown and the other where the true value is said to be theoretically known. As an example of the first case, let us consider the process of measuring the line AB with, let us say, an engineer's scale graduated to 0.01". Ten measure-



ments of one such line in this manner gave the data in Table 10.

4.000	3.996	3.996	3.990	3.994
3.996	3.994	3.994	3.992	3.992

Table 10

Obviously the procedure of measuring the line in this way could at least theoretically be repeated again and again giving an infinite sequence of numbers:

$$X_1, X_2, \dots, X_i, \dots, X_n, X_{n+1}, X_{n+2}, \dots, X_{n+i}, \dots \quad (25)$$

As an example of a case when the true value is said to be known theoretically, let  $\alpha, \beta, \gamma$  be the three angles of any triangle measured in degrees. Let  $X$  represent the sum of these angles. Then the true value  $X'$  of  $X$  is said to be  $180^\circ$ . In practice, however, if  $\alpha, \beta, \gamma$  are measured with, let us say, a protractor, then the observed sum  $X$  will in general be found to differ from the true value  $X'$ . An actual experiment involving the measurement of the three angles of ten triangles gave the following results:

(1)	(2)	(3)	(4)	(5)
<u>Triangle</u>	<u>a</u>	<u>b</u>	<u>γ</u>	<u>α + β + γ</u>
1	14.3°	15.1°	151.0°	180.4°
2	51.1	54.9	74.0	180.0
3	18.5	16.6	145.0	180.1
4	48.0	60.2	72.0	180.2
5	51.5	36.5	92.5	180.5
6	77.0	45.2	58.3	180.5
7	9.5	29.6	140.6	179.7
8	24.5	35.1	120.4	180.0
9	14.4	152.5	13.0	179.9
10	16.2	147.3	16.3	179.8

Table 11

The numbers in column five reading from top to bottom constitute a sequence such as (25).

Let us consider the concept of accuracy in the case where the true value  $X'$  is given theoretically. Let us lay off on a line a number representing this true value Fig. 29. Then let us locate on this line two other points  $X = X' - l_1$  and  $X = X' + l_2$ .

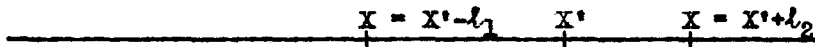


Fig. 29

Accuracy with respect to the range  $X' - l_1$  to  $X' + l_2$  has to do with the clustering of the numbers in a sequence of type (25) within this range.

Before considering further the concept of accuracy in this example let us consider the corresponding concept of precision. Let  $\bar{X}'$  be the average of the numbers constituting the sequence. Locate this average and two other points  $X = \bar{X}' - l_1$  and  $X = \bar{X}' + l_2$  on a line Fig. 30. Let us assume a case where the theoretical true value  $X'$  is not equal to  $\bar{X}'$ .



Fig. 30

Precision has to do with the clustering of the numbers in a sequence of measurements within the range  $\bar{X}' - l_1$  to  $\bar{X}' + l_2$ .

If we view accuracy and precision in this way, it is obvious that for a sequence in which

$$X' = \bar{X}', \tag{26}$$

or where the theoretical value  $X'$  is the same as the average  $\bar{X}'$  of the sequence, the concepts of accuracy and precision become the same if the method of measuring the clustering effect is made the same in the two cases. The customary theory of errors measures the clustering in the same way and in effect assumes that equation 26 is satisfied. Hence it is that the two terms are often fused into one and the same meaning.

Let us for the moment take as a measure of clustering the fraction  $p'$  of the numbers in a sequence falling within the chosen range. Then we might conceive of comparing two sequences of measurements of the same true value  $X'$  in respect to accuracy corresponding to a chosen interval by comparing the corresponding fractions of numbers falling within this range. In the case where  $l = l_1$  is chosen equal to  $l_2$ , it is sometimes convenient to speak of the accuracy for a chosen value  $p'$  as the ratio

$$\text{percent accuracy} = \frac{l}{X'} \times 100. \tag{27}$$

Such an accuracy expressed as a percentage, however, is obviously dependent upon the value of  $p'$  chosen. Presumably we may define precision in much the same way as

$$\text{percent precision} = \frac{l}{\bar{X}'} \times 100, \tag{28}$$

where the precision as thus given corresponds to a chosen value of  $p'$ . Equations (27) and (28) constitute the basis for defining accuracy and precision in percentage. It should be noted that if the true value  $X'$  is equal to the expected value  $\bar{X}'$ , then the percentage measures become the same.<sup>1</sup>

-----  
1. Of course it is sometimes assumed in the literature of the theory of errors that not only does  $X' = \bar{X}'$  but also that the distribution in the sequence is random and follows the normal law of error with standard deviation  $\sigma'$ . Then the probable error which is  $.6745 \sigma'$  is taken as a measure of accuracy and  $h' = \frac{1}{\sigma' \sqrt{2}}$  is taken as a measure of precision,  $\sigma'$  being the standard

deviation of the law of error. Here again perhaps we find another reason why the two concepts are often taken as synonymous in the literature, because they are both expressible in terms of the standard deviation  $\sigma'$ .

SOME CRITICAL COMMENTS

We started out to find what it means in a verifiable manner to specify that the accuracy shall be 1% and the precision shall be 1%. Let us try out the concepts of accuracy and precision as defined by equations (27) and (28). In order to verify the requirement as to accuracy, we must be able to find  $l$  and  $X'$  and must be given the value  $p'$ . Likewise for precision we must be able to find  $l$  and  $\bar{X}'$  and must be given  $p'$ . In so far as such concepts are used in practice, however, there is more tacitly involved than is explicitly stated. For example, the fraction  $p'$  of the numbers in a sequence between any two arbitrarily chosen limits is referred to as a probability with the implication that the sequence is one to which the concept of probability may be applied, or in other words, that the sequence is a "random" one. The assumption that the meaning of such a requirement for accuracy is operationally verifiable at least in a theoretical sense involves the following specific assumptions:

- $a_1$ . That the value  $p'$  is given.
- $a_2$ . That the value  $l$  can be found.
- $a_3$ . That it is possible to set up an operationally definite criterion of randomness.
- $a_4$ . That the value  $X'$  can be found.

Likewise, for the case of precision the following assumptions are involved:

- $b_1$ . Same as  $a_1$ .
- $b_2$ . Same as  $a_2$ .
- $b_3$ . Same as  $a_3$ .
- $b_4$ . That the value  $\bar{X}'$  can be found.

In respect to the assumption  $a_1$ , there is no difficulty because it involves an arbitrary choice. Given an infinite sequence there is no theoretical difficulty in finding a value of  $l$  such that the range  $X' \pm l$  will include the fraction  $1-p'$  of the numbers in the sequence, provided of course that  $X'$  be known.

Let us now look at assumption  $a_3$ . In what sense is there an operationally definite criterion such that if a given sequence meets this criterion, we can say with certainty that the sequence is random? We considered this question in the first chapter when discussing means of characterizing the

sequence produced under conditions of a state of statistical control. It was pointed out that there is an indefinitely large number of criteria for such a sequence, most of which are not even known at the present time because every new development in statistical distribution theory adds to the possibilities. This situation constitutes a theoretical difficulty. There is, however, another and perhaps even more fundamental difficulty here involved. Let us assume, for example, that such a criterion could be found, and that we had an infinite sequence before us. The fact that this sequence satisfies the criterion does not mean that, if the numbers in the sequence were written on chips, these chips were thoroughly mixed and then drawn one at a time without replacement thus forming a new sequence, this new sequence will also necessarily satisfy the criterion. Nevertheless the concept of random implies that this second sequence is a random sequence of the same set of numbers as the first.

Now coming to the fourth assumption  $a_4$ , I do not know any operation by which we can even theoretically find the true value  $X'$ . I think this will be admitted for any measurement such as that of the length of a line AB. However, someone may point out that in certain problems such as the measurement of the sum of the angles in a triangle, there appears to be a theoretical value. We should keep in mind in this case that the claim that the sum of the angles of a triangle is  $180^\circ$  rests upon the choice of a particular set of postulates. For another well-known set of postulates the sum is theoretically greater than  $180^\circ$  and for still another set, the sum is theoretically less than  $180^\circ$ . Hence there are inherent theoretical difficulties in the fourth assumption.

Passing to the case of precision, we have the same theoretical difficulty in the third assumption  $b_3$  as we did in  $a_3$  because the assumptions are the same. Theoretically, however, I do not see any difficulty with assumption  $b_4$ , because there is nothing to hinder one from taking more and more terms in a given sequence. True it is that we could never reach the end of the sequence but we at least can visualize a process of approaching the end of the sequence whereas, in the case of trying to find a true value  $X'$  one does not even know that he is on the right track.

Obviously there are insurmountable practical difficulties in all except the assumption that  $p'$  is given. For example,  $a_2$  and  $b_4$  although theoretically verifiable are not verifiable from a practical viewpoint. It would take a whole flock of Methuselah's to count even one infinite sequence!

In the light of such considerations, we must conclude that there are some formidable theoretical as well as practical difficulties in trying to use the concepts of percentage accuracy and precision expressed in equations (27) and (28), if we are to require that the meaning is to be operationally definite. If we were to consider the problem of making definite statements involving the use of these concepts, we would run into still more complications because such statements must be probable inferences. Let us, therefore, approach the problem from a different angle.

#### OPERATIONALLY VERIFIABLE CONCEPTS OF ACCURACY AND PRECISION

Introductory Comments - In the example of measuring the line AB as considered above, we may think of the process of measurement as an operation generating the sequence (25). There are two aspects of such an operation that are fundamental for our present purposes: a) one is the actual physical operation of measuring employed by an observer, and b) the other is the potentially infinite sequence. If we are to find any quantitative measure of accuracy and precision, it must of course be in terms of chosen characteristics of such a sequence.

Let us think for a moment in terms of this sequence. Customary theory assumes that it is random. However, we must have some practical and operationally verifiable criterion of randomness in the sense considered in the first chapter. Fundamentally this means that a sequence must satisfy certain criteria of statistical control in order that we may be justified in assuming the existence of a constant probability as is done in customary practice and in order that we may make valid predictions based upon a sample in the same way that we would base predictions on a sample drawn from a bowl-universe. We have seen, however, that only a few, if any, sequences usually assumed to be random satisfy such criteria. It follows that, in so far as randomness can be interpreted operationally in terms of criteria of statistical control, it seems that the assumption of randomness in customary error theory is seldom valid.

Hence customary measures of accuracy and precision based upon the assumption of randomness or the existence of a statistical state of control are seldom justified in practice.

Even under conditions where the operation of measurement is not in a state of statistical control, we need to have some information about the reproducibility and accuracy of the method. From the viewpoint of reproducibility, about as far as we can hope to go in the general case is to establish a tolerance range in the sense discussed in the second chapter. That is we must be satisfied with trying to set up a range

$$X = L_1 \text{ to } X = L_2$$

such that the probability  $p$  of falling outside this range does not exceed some specified value  $p'$ . We shall later return to see just what this statement means operationally.

Next, however, let us consider for a moment the matter of accuracy. Let us look again at the simple case of measuring the length of the line AB with an engineers scale. Associated with the use of each such scale there is potentially a sequence such as (25). Likewise there are many other potential sequences corresponding to other methods or operations of measurement. Now it is obvious that if the line AB has a true length  $X'$  and if different methods of measuring actually measure this length, then there should be some kind of consistency between the results obtained by the different methods. In this way the assumption of a true value more or less naturally leads us to impose the requirement of consistency on the results obtained by different methods of measuring and in this way, we come to think of accuracy in terms of certain observable aspects of consistency.

All that has been done thus far in this section is to point out that an operational approach leads one to think of precision in terms of a tolerance range and to think of accuracy in terms of consistency. Much remains to be done in order to attain operationally definite criteria for these two concepts.

Measurement from the Operational Viewpoint - It is significant for what follows that there are two aspects to an operation of measurement, one which is quantitative in character and one which is qualitative and represents the condition associated with the quantitative aspect. A part of the qualitative



condition C is physical in character and a part is "human" in so far as it depends upon the operator. Schematically for each X arising in an observable sequence, we have:

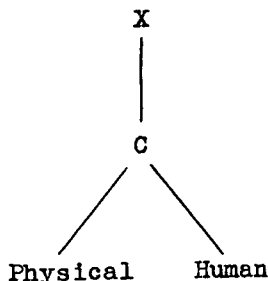


Fig. 31

To illustrate a little more definitely the difference between the two components of C, we may think of the potential sequences of type (25) corresponding to the use of different engineers scales by the same operator in measuring the length of a line, or of the set of sequences obtained by a group of operators using the same scale. The first point I wish to make is that in trying to fix upon operationally verifiable criteria we must in the end take into account the two aspects of an operation as indicated in Fig. 31.

Next let us consider simply the potential sequence of X's associated with any method of measuring. In the first place, we should note that, since such a sequence is potentially infinite, we cannot have a measure of either accuracy or precision that can be verified in practice and at the same time involve a characteristic of the whole sequence. On the other hand, no matter how many measurements one has made in the past by a given method - in other words, no matter how large a sample  $n$  of an infinite sequence one has examined the part of the sequence of interest is almost always a part not yet observed. Thus there are the three parts of a sequence shown in Fig. 32 that call for attention.

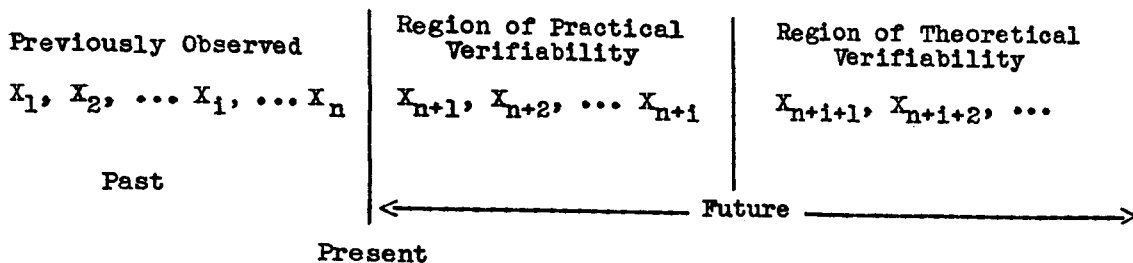


Fig. 32

Hence if one is to characterize accuracy and precision of a given sequence in a way that can be checked experimentally and in a finite time, such characterization must apply to the central portion of the sequence as shown in Fig. 32. For example, this rules out from the viewpoint of practical verifiability, statements involving the concept of a statistical limit. In fact, all estimates applying to characteristics of the infinite sequence are outside the region of practical verifiability.

Now we are in a position to say something definite about requirements appearing in specifications as to accuracy and precision and also statements about accuracy and precision, such as we considered above in the introduction. As to the requirements, they must in order to be practically verifiable involve only the central portion of the sequence, Fig. 32. Likewise statements involving accuracy and precision are of the nature of predictions which can only involve this same portion of the sequence if they are to be verifiable in a practical sense. Such predictions, of course, may or may not be valid and our degree of belief in their validity will depend in part upon the evidence provided by the sample of  $n$ .

Next let us consider a little further the subject of consistency. In this case, as we have already seen, we have for consideration a set of potential sequences instead of just one, which may be symbolically represented as follows:

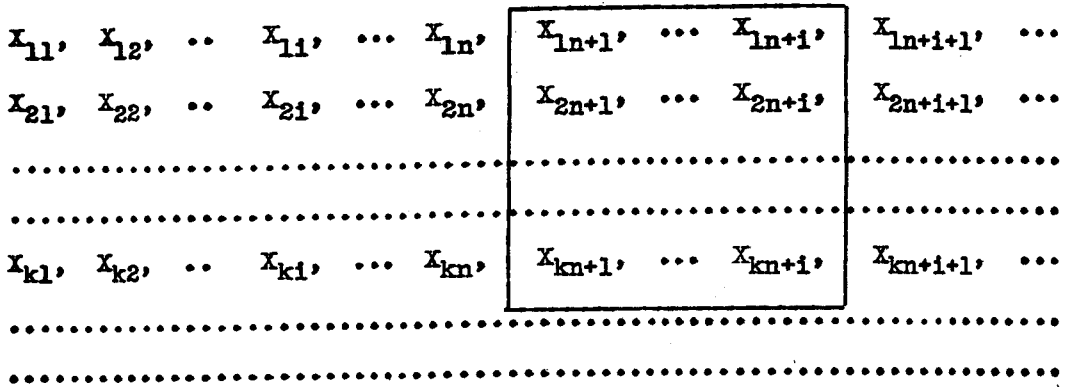


Fig. 33

17836

Operationally each sequence may be thought of as the one shown in Fig. 32. In this case, however, each sequence not only is infinite, but there are potentially an infinite number of different sequences. To make the meaning of accuracy operationally verifiable in a practical as contrasted with a theoretical sense, the criteria of consistency must be limited to the set of numbers in the enclosed area of Fig. 33, it being kept in mind that the values of  $i$  and  $k$  may be chosen at will. For example, one method of measurement may be chosen as a standard of comparison and hence in Fig. 33 there would be only the standard sequence and the one to be compared with the standard.

Now we are in a place to note an essential characteristic of any operational criterion that is practically verifiable. Its use must by definition lead to a two-fold classification - yes or no - of any sequence. Hence for reasons which we have previously considered, its use cannot serve to verify a probability. For example, let us consider the concept of a tolerance requirement on the single observations in a sequence such as (25). We conceive in this case of two values  $X = L_1$  and  $X = L_2$  such that the probability of falling outside the interval  $L_1$  to  $L_2$  is less than some specified value  $p'$ . Now, if we apply such limits to an observable part of a sequence, we can only observe that the fraction of the numbers in this portion of the sequence falling outside of the tolerance limits is either less than, equal to, or greater than  $p'$ . Obviously this fact does not verify any probability statement.

It is of interest to note that we may think of tolerance limits on all the numbers within the enclosed area of Fig. 33 instead of applying the concept to numbers within the region of practical verifiability in a single sequence.

In fact, it is much this concept that is often applied in discussing errors of measurement of some physical constant, except that in this case the tolerance limits are set on the averages of the parts of the separate sequences shown in Fig. 33 within the enclosed area.

Criteria for both accuracy and precision may be expressed in terms of limits on functions of the single observations instead of expressing them in terms of the single observations themselves. For example, in the case of accuracy, we may compare the operationally verifiable portions of the sequences for averages, variances or any one of many other functions. If  $\psi$  be any such function, the criterion to be verified by looking at the part of the sequences within the enclosed area, Fig. 33, is of the nature

$$\psi_1 \leq \psi \leq \psi_2 \quad (29)$$

where  $\psi_1$  and  $\psi_2$  are two previously assigned tolerance limits.

From the viewpoint of statistical theory, it is significant that the concepts of accuracy and precision lead to the use of tolerance limits and not fiducial limits, when we try to set up operationally verifiable criteria.

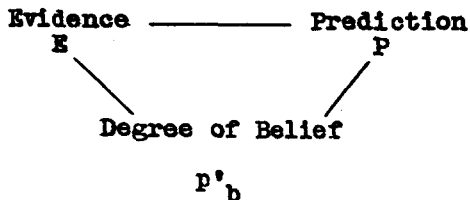
Specification of Accuracy and Precision - The question naturally arises as to how the limits  $\psi_1$  and  $\psi_2$  shall be set. If the specification is to be verifiable in the practical sense, the first step is to specify the portion of the sequence (32) or of the sequences (33) that is to be taken in the process of verification. The second step is to specify the function  $\psi$  and the limits  $\psi_1$  and  $\psi_2$ . Theoretically, of course, the tolerance limits  $\psi_1$  and  $\psi_2$  may be set at will. However, as pointed out in the last chapter it is necessary from a practical viewpoint to insure that it is feasible to meet them. Particularly in the case of measuring devices, the limits  $\psi_1$  and  $\psi_2$  may be expressed as a percentage of the average of the numbers to be taken in the process of verification. Or again one may specify that the ratio of the standard deviation (or some other measure of dispersion) of the set of numbers obtained in the process of verification to the average of these numbers shall not be greater than some specified value. It should be noted, however, that such a method of specifying a tolerance range is not so desirable when there are assignable causes of variation present as when these have been removed. Tolerance limits

may, however, be specified by other methods irrespective of whether or not assignable causes are present. If, however, it is desirable to try to attain a statistically controlled or random state of variability, it may be desirable to specify an operationally verifiable criterion of control. The control criterion as thus used is like a tolerance range in that it constitutes simply a basis for classification. In this case it may be desirable to take into account the fact that all members of the same sequence may not be taken under what even appear as the same essential conditions and the specification may require a special grouping to take advantage of this information.

JUDGMENT OF ACCURACY AND PRECISION

Let us consider the statement: The accuracy (or precision) of this test method is 1%. Such a statement is of the nature of a prediction resting on previous evidence. The degree of rational belief  $p_b$  (or assurance) that one may place in a prediction P inheres in its relation to the available evidence E. If a prediction is expressed in terms of an operationally verifiable requirement such as that considered in the previous section, then its validity may be checked at will.

It should be noted, however, that the process of verifying a prediction in the sense just indicated does not verify the statement in which the prediction occurs. The sentence or statement that the accuracy (or precision) of any experimental procedure is such and such must be viewed from the three aspects of knowledge considered in an earlier chapter. That is we must consider the diagram:



where  $p'_b$  is a so-called objective degree of belief.

Scientific statements are always probable only as contrasted with statements in mathematics and logic that are of a deductive character and hence are certain. Deductive statements are either right or wrong and may be verified once and for all by using the agreed-upon formal rules. For example, let

us consider a statement about mathematical accuracy.

In elementary mathematics we learn that if we let  $y$  be the circumference of a circle and  $r$  the radius, then  $y = 2\pi r$ . Later we are given ways by which we may calculate  $\pi$  to as many decimal places as we wish. In this way we find that<sup>1</sup>

$$\pi = 3.14159\ 26535\ 98793\ 23846\ 26433\ 83280$$

to 30 decimal places. Of course, these 30 places do not give us the exact value of  $\pi$  - the exact value is transcendental and cannot be written down. We know, however, that it lies between the value given above and that value plus  $10^{-30}$ . In other words, we know that the maximum possible error we could make by taking  $\pi$  as given by Eq. 6 would be less than

$$\frac{1}{1000000000000000000000000000000} \\ |\pi - \text{true value of } \pi| < 10^{-30}$$

Put in still another way we can say that  $\pi$  as given above is accurate within less than one part in  $10^{30}$ . Every time that  $\pi$  is calculated to 30 or more decimal places we can be certain that, barring mistakes, the first 30 decimal places will be those given above. That statement about accuracy is true for all time, so long as we accept the rules of computation.

Let us now contrast with this statement about mathematical accuracy, a statement about the accuracy of the measurement of the sum of the angles of a triangle based upon the evidence provided by the ten observations in Table 11 and assuming that  $180^\circ$  is the theoretically true value. Assume that we wish to state the accuracy in terms of a tolerance range on single observations that will include, let us say, 99% of the next 1000 measurements by this same method. Let us assume that the percentage accuracy in this case is to be stated as  $\frac{\Delta\theta}{180^\circ}$ . Where  $\Delta\theta$  is to be derived from the data, and  $180^\circ \pm \Delta\theta$  is

1. cf. Pearson, Karl, Tables for Statisticians and Biometricians, Part II, page 262, published by University of London, 1932. For one method of evaluating  $\pi$  see Theory and Application of Infinite Series, K. Knopp, English translation, Blackie and Son, Ltd., London, 1928, pp.252-254. The number  $\pi$  has been computed to 700 decimal places.

the tolerance range criterion for accuracy. The concept that there is an objective degree of rational belief  $p'_b$  is of importance in much the same way that the concept of true value has been shown to be important above. The concept of  $p'_b$  is, as I see it, translatable into the concept that there is a "best" way of setting up the value  $\Delta\theta$ , given the ten observed values in Table 10. The process of verifying a given method of setting up a  $\Delta\theta$  would consist in testing out different methods of obtaining a  $\Delta\theta$  from samples of ten observations and not by the means which would be used in finding out if a particular value of  $\Delta\theta$  derived from this sample of ten is justified. This fact is of great practical importance in that it shows why anyone making statements about accuracy and precision upon the basis of given evidence is responsible for trying to make the best possible statements. In the light of the discussion in the second chapter about setting tolerance ranges, it is evident that in trying to make the best prediction we must first apply some practical operation for testing for control. If evidence of control is found, the problem of finding the best method is statistical in character, but if evidence of assignable causes is found, the best method is more involved.

Importance of Quantity and Kind of Evidence - The fact that accuracy and precision from an operationally verifiable viewpoint is of the nature of a tolerance instead of a fiducial limit is of outstanding importance in that even under ideal conditions of drawing from a normal distribution in a bowl, the validity of a prediction depends upon the number of observations. Whereas statements involving predictions in terms of fiducial limits are, under such ideal conditions, just as valid for small samples as for large ones, this state of affairs does not hold for predictions in terms of tolerance limits.

Furthermore, in trying to judge the validity of a statement about accuracy and precision, we must not only consider the quantity of information but also the available evidence as to how many assignable causes of variation have previously been found and eliminated. Finally, we must try to weigh the human factor entering through the condition C constituting the qualitative aspect of an operation, Fig. 32. Hence it is that the judge of accuracy and precision must always present his evidence in making statements involving these terms.

Why Accuracy and Precision Should not be Confused - It would seem trite to stress the significance of the difference between accuracy and precision if these were not so thoroughly confused in the literature. For example, let us consider the simple case of trying to make a statement involving these two concepts and based upon the ten measurements of the length of a line, Table 9, by one method. So far as I see, one could say something with justification about precision but not about accuracy.

In general, reproducibility is as much different from consistency as day from night. The nature of the data that must be available in order to make valid predictions for precision is not the same as that for accuracy, and the methods of verifying predictions in the two cases are different. Likewise, the best method of making a prediction in terms of precision may under certain conditions become purely statistical in character whereas this is never true for accuracy. In fact, we can under ideal conditions approach closer and closer to the "objective" value of precision simply by the process of repetition whereas we can perhaps never hope to discover any simple road like this to follow in approaching accuracy, because we must always depend so largely upon human ingenuity in discovering different methods to be used for comparison purposes.

Specification in Relation to Judgment - It is perhaps desirable to stress the intimate relationship between the method of stating the requirements as to accuracy and precision and the effort in time and money that will be required in obtaining a satisfactory degree of assurance that apparatus and mechanisms passed in the process of inspection will live up to the requirements. For this reason if no other, it is desirable for the specification requirements to be stated in experimentally verifiable and definite form. On the other hand, it is difficult to do this once and for all, because the requirements in order to be satisfactory from an economic viewpoint must be such as can be met without undue cost. Under such conditions it may in certain instances be desirable to leave the specification in simply a theoretically verifiable form. It is for this reason that specification and inspection engineers find it necessary to cooperate in their efforts to attain accuracy and precision. Progress in this direction should materially increase with a broader understanding of the operationally verifiable significance of these terms and the relation of pre-



dictions in these terms to the process of providing adequate inspection.

### SOME CONCLUDING REMARKS IN RESPECT TO STATISTICS

Our brief excursion started with a consideration of the state of confusion in which the terms accuracy and precision stand in the literature today. Though definitions of these terms in classic theory make them different in that the first involves the concepts of true value  $X'$  and the second involves the concept of expected value  $\bar{X}'$ , this difference is often confused in that later in the development of the theory it is assumed that  $X' = \bar{X}'$ . The first of these terms is operationally verifiable in the theoretical but not in the practical sense. The second term is not even theoretically verifiable until we translate it into the concept of consistency which permits of an arbitrary choice of operationally verifiable tests for consistency, though complete verifiability is still not attainable. In all cases, the customary meanings of accuracy and precision are made to rest on the condition that a state of statistical control exists and this can be translated into an arbitrary choice of criteria for control which practice has shown to function satisfactorily. Customary theory leads to the use of fiducial limits which are not operationally verifiable where the true value  $X'$  is unknown. To obtain operational verifiability, it becomes necessary to introduce a tolerance limit type of criterion, the validity of which depends upon the weight of information (as, for example, sample size) in a way entirely different than is the case for fiducial limits. In this way we come to see the importance of large as contrasted with small samples. Finally, we have seen that since, in general, the criteria for statistical control are not satisfied in practice, we are seldom in a place to use customary theory which is based upon the assumption of randomness.

One of the most important points brought out in the discussion is the inherent difference from the viewpoint of operational verification between a specified requirement for accuracy and/or precision and a statement involving these terms. The former is subject to verification. The second involves the concept of a prediction resting upon stated evidence through a degree of belief type of probability. The concept of an objective degree of rational belief  $p'_b$  is translatable into an operational technique for discovering the best prediction in terms of tolerance limits upon the basis of given evidence. In so

far as verifiability for a statement is attainable it includes the technique of verifying a specificational requirement and an additional technique for comparing methods of setting the tolerance limits.

What rôle does statistical theory play? It can with the introduction of operationally verifiable concepts be made to play a very important rôle. In this form it furnishes us the mathematical distribution theory required for control criteria, and powerful tests for judging the state of consistency including application of tests for significant differences, and the technique of the analysis of variance. Under controlled conditions the problem of setting criteria for precision is purely statistical and that of setting criteria for accuracy is at least partly statistical. In fact, statistical theory as we have seen furnishes us the background for the development of all of the practical techniques to be used although under normal conditions this theory must be supplemented by other techniques which do not depend for their validity upon improving inference through repetition.

*Mrs Nelson*

