NATIONAL BUREAU OF STANDARDS REPORT

NBS Project		NBS Report
3011-60-0002	26 March 1952	1539

STATISTICAL UNITS OF MEASUREMENT

by

W. J. Youden

NBS

U. S. DEPARTMENT OF COMMERCE
NATIONAL BUREAU OF STANDARDS

FOREWORD

The adoption of statistical procedures for the interpretation of data will be encouraged, first, by drawing the attention of scientists and engineers to statistical procedures that are the counterpart of the questions which inevitably arise in the examination of data, and, second, by indicating that these procedures are valid for small sets of data.

The paper was prepared as the first of a series of four lectures sponsored by the Philadelphia Chapter of the American Society for Metals.

J. H. Curtiss
Chief, National Applied
Mathematics Laboratories

A. V. Astin
Acting Director
National Bureau of Standards
26 March 1952

STATISTICAL UNITS OF MEASUREMENT*

by

W. J. Youden**

Engineers are accustomed to devise suitable measuring units to express the magnitude of the physical properties of engineering materials. In many cases these units are simple combinations of the basic units employed in everyday life. The tensile strength of an aluminum wire may be 35,000 pounds per square inch. A pound per square inch is a unit of measurement for tensile strength and the number 35,000 expresses the engineer's verdict regarding this physical property of the metal. Such units are so familiar to us that they are usually taken for granted, especially when the unit can be visualized and found meaningful.

There are many cases in which the unit or scale has an arbitrary basis. The scale of hardness for minerals is an example. The values 1 to 10 are assigned to ten materials ranging from talc to diamond. Other materials are rated by their response to scratching with specimens chosen from the standard list. Obviously a longer list of standard materials would give finer divisions. This scale of hardness is something which men have agreed to use to convey their knowledge of the properties of materials. Other hardness scales are available for metals, such as the diameter of the indentation made by a hardened steel ball under specified circumstances.

Both these scales of hardness have a recognizable meaning in terms of everyday experience with the concept known as hardness. Sometimes the connection is more remote. If the measurement process is destructive recourse may be had to some other form of measurement the results of which are closely correlated with the performance of the material. It is difficult to see any simple property that is measured by the Charpy Notched-Bar Test yet such measurements are considered, at least by some, to furnish information on certain performance characteristics of steel.

Making measurements of one kind or another is a primary activity for most engineers and scientists. Attention is directed to the answers, that is, the magnitudes which are obtained. Judgments

---

*Talk given before the Philadelphia Chapter of the American Society for Metals, March 4, 1952.

**National Bureau of Standards, Washington 25, D. C.

are then formed by appraising these magnitudes against the individual's accumulated experience with such measurements. Often the result or measurement is directly compared with some specification to determine the fate of the material in question. Men have learned also by experience that measurements vary and are aware that the fate of border-line material may depend on this variation. On one test the material may fall just short of specification. On a repeat test the inherent variation in the measuring process may turn up a result which meets the specification. The material has not changed; the uncertainty lies in the limitations of the procedure for making the measurement or possibly in the sample selected or in both.

In the early stages of the development of a craft attention is properly directed to refining the measurement procedure. Attempts are made to reduce the variation shown by repeated measurements on the same material so that it can be neglected. When this is achieved there is no uncertainty about whether or not a product meets specification or about the performance of the product. This happy state of affairs is not easy to attain. Sometimes, by building extra quality into the product, one can be sure that all measurements meet specifications and there will be no rejections. This can be expensive. Whenever we crowd the specification we risk the chance that good material will be rejected simply because measurements vary. Consequently great efforts have been made to improve measurements and, in recent times, to making the most of the measurements we have. It is at this point that someone begins to look into the subject of statistics since it is rumored that statisticians can get information out of data.

It will be my concern to give some indication of what it is the statistician has to offer. The somewhat lengthy introductory remarks on units had a purpose. We shall need a unit to measure variation. I shall mention three common units that are used to give quantitative expression to our knowledge of the variation that measurements may exhibit.

Even if but two measurements are available the difference between them is a measure of the variation shown by such measurements. Once sufficient experience with a given procedure has been accumulated it is possible to pass judgment on the agreement shown by a pair of readings and determine whether or not confidence may be placed in the average of these two readings. If more than two readings are available a simple extension of the rule of taking differences leads to using the range, i.e., the difference between the largest and smallest measurements in the set, as a measure of the variation. One important drawback to this simple unit is that it needs adjustment depending upon the number of measurements in the set. Five measurements will, on the average, have a range more than twice as large as the average difference between two measurements. The average difference between the largest and smallest of a group of 14 measurements exceeds three times the average difference between a pair of measurements. The intrinsic variation of the measurement operation concerned is the same regardless of the number of measurements taken, so a unit that depends upon the number in the group must be adjusted by an appropriate factor.

Another unit for measuring variation is the average deviation -- simply the average of all the differences (without regard to sign) obtained by subtracting the average for the group from every measurement. This unit is in widespread use by engineers and scientists. It would not be used so much if the user knew that this unit, like the range, depends upon the number of measurements in the group. Indeed the average deviation, if groups of three measurements are taken, is, in the long run, but seven tenths the correct value. The correction factor depends on the number in the group. Almost invariably this correction is ignored. No engineer would use a unit of length that got larger for long objects than for short ones.

Statisticians have their favorite unit for measuring variation among measurements. It is called the standard deviation and is computed by squaring the differences used in the average deviation and dividing the sum of the squares by one less than the number of measurements. The final step of taking the square root returns the unit to the same scale as the original measurements. This seems like a tedious process but so is the process of taking the measurements in the first place. The square of this unit meets the first requirement that, on the average its calculated size is not influenced by the number of measurements used in its computation. Equally important, this unit is particularly well adapted to answering the questions that experimenters are bound to ask about their data. The numerical operations with this unit are simple. Furthermore, the standard deviation extracts more information from the data than the range or average deviation.

A word of caution may be interjected at this point. We are here considering measurements that are independent and of equal precision. If the measurements are not independent the information furnished by a measurement is, in part, a duplication of informatio already supplied by other measurements. This reduces the effective size of the set of measurements. Also, if measurements have varying precision it becomes necessary to weigh them appropriately otherwise a single imprecise reading may dominate the others.

Suppose we enumerate some of the questions that experimenters ask about data and give some examples to show how the standard deviation is put to work to answer their questions. It will not be possible to do a complete job in this paper but some examples will show that a unit for the measurement of variation is indispensable for a really close examination of a set of measurements.

In the first place it is often desirable to be able to state the variability of a measurement procedure.

TABLE I

	Measurements	Diff. from Average		Diff. Squared
	2.39	.06		.0036
	2.31	-.02		.0004
	2.38	.05		.0025
	2.32	-.01		.0001
	2.27	-.06		.0036
Average	2.33		Sum	0.0102

$$\text{Standard Deviation} = \sqrt{\frac{0.0102}{5-1}} = \sqrt{0.0025} = 0.05$$

Table I shows the computation or estimation of the standard deviation from a group of five measurements. The result, 0.05, is called an <u>estimate</u> of the standard deviation because, obviously enough, another set of five similar measurements will lead to another result differing more or less from the first. The original readings, or estimates of the property, show variation and consequently we may anticipate variation among a series of estimates of the variation.

Further questions immediately come to mind. Suppose these measurements have been obtained by a modification of a test procedure which long experience has shown to have a standard deviation of 0.12. Are these data sufficient to warrant the conclusion that a real improvement has been achieved in the test procedure? The statistician will divide the sum of squares, 0.0102, from above by the square of 0.12 obtaining

$$\frac{0.0102}{0.0144} = 0.71 \quad .$$

This ratio, known as Chi square, is looked up in a statistical table (see Appendix 1) which informs the statistician that only once in 20 times would he get five measurements to agree as well as these do if the procedure has not been improved and it still has the original standard deviation of 0.12. Perhaps this is the lucky one in twenty shot. The usual conclusion, however, is that an improvement has been effected.

There is another question we might consider even before the measurements are taken. We may set our sights a bit lower and decide, in advance, that we would be pleased if the modification in the method cut the actual standard deviation (not merely its estimate) to one half the former figure of 0.12. If the modification has really brought about this much improvement we would like very much to establish that fact. It is necessary to face the situation that, in any given group of measurements, the estimate of the standard deviation may, by chance, be low and we could be misled. We shall insist, therefore, that the estimate obtained from the measurements, when they become available, be below such chance values or we will not be impressed by it. On the other hand, the modification may really cut the standard deviation down to one half its former value and nevertheless the set of measurements obtained give, by chance, an estimate somewhat above the value 0.06. How many measurements should we take to prevent our concluding an improvement has been achieved when in fact it has not and also give us a good liklihood of catching this improvement if it really has been achieved. Tables have been prepared to guide the experimenter in this matter. If the experimenter decided he wants to risk only one chance in twenty of believing in an improvement, when none has taken place, and to have nine chances in ten of picking it up if this 50 percent reduction has been achieved, such tables show that 12 measurements should be collected. If this seems a large number, it means that some people underestimate the amount of work that it takes to demonstrate that such an improvement has been made. A short table is given in Appendix II.

These techniques for assessing the variation among measurements are, of course, equally applicable to the assessment of the variation shown by units of a product that is being manufactured. Almost always the objective is to achieve greater uniformity in a product and various steps are taken with this in mind. Sooner or later the question arises: Has the variation been reduced by the steps taken? It may be that there is no limit on the number of units available. On the other hand, if the tests are destructive of a valuable item, or if the measurement of quality or performance of the product is expensive, then it is useful to be able to make an estimate, in advance, of the amount of work required in order to arrive at a sound decision. Not the least benefit accruing from the statistical approach is that it removes the matter from the field of personal opinion and submits it to an objective criterion.

Another variation of this question is often encountered. A purchaser may be offered items from two suppliers. Naturally he will wish to choose the more uniform material; the one showing least variation. Presumably this will assist him in producing a uniform product. Here the variability of neither material is known in advance. The purchaser must first estimate the standard deviations. These are compared by taking the ratio of the squares of the standard deviations, setting up the fraction so that the ratio is greater than unity. This ratio, called F, must attain a certain critical value before it can be concluded that a difference in variability between the products really exists (see Appendix III). Furthermore, statistical tables are available to guide the purchaser in the amount of data required to be reasonably sure of detecting a difference in variability that would be important in his process. (See Appendix IV).

These examples by no means exhaust the list of ways of making good use of knowledge of the variation exhibited by a group of numerical quantities. This variation is one of the reasons for making repeat tests. It is well known that the average of a serie[s] is less prone to variation than the individual measurements enteri[ng] into the average. In fact, the standard deviation of the <u>average</u> of <u>n</u> measurements is estimated by dividing the standard deviation (as calculated in Table I) by the $\sqrt{n}$. This is one step toward answering a question of greater interest to the man with the data. He would like to know whether an interval can be specified such that, when centered on the observed average, he can have a certain confidence that it will bracket an average based upon a great many measurements. If this interval does not overlap some rejection value the data do not contradict the claim that the product meets the specification. Of course more data might show that the product does not meet the specification.

Here again there are two distinct situations -- first, the case where the standard deviation has been established through an accumulation of earlier records and, second, the case in which the limited set under examination is the sole source of information on the variation of the data. It is reasonable to expect that a narrower interval can be set if there is available a well established figure for the standard deviation. Indeed, for 95 percent confidence it is customary to make the interval extend two (more accurately 1.96) standard deviation* on either side of the average. The factor becomes larger as the number of measurements available for estimating the standard deviation decreases. With 20 measurements the factor is 2.09, for 10 measurements 2.26, and for 5 measurements 2.78. These are the well known _t_ values taken from widely available Tables of _t_ (See Appendix V). The larger factors reflect the additional uncertainty introduced by the variation in the estimates of the standard deviation which naturally enough gets worse as the number of measurements becomes smaller.

When earlier records provide a reliable estimate of the standard deviation it is possible to specify the number of measurements required to have a given chance of picking up a shift in the average. The shift is expressed in units of the standard deviation and a table may be consulted (see Appendix VI) for the required number of measurements. The table vividly shows the large number of measurements necessary to be reasonably sure of catching a small shift in the average.

There are still some workers who hold that statistical procedures can only be used when large masses of data are involved. This simply is not true. Often these same workers do not hesitate to express their intuitive judgments upon rather small sets of data. But statistical procedures are nothing more than the translation of such intuitive judgments into more exact appraisals by making use of an appropriate unit of measurement for the evaluation of the variation in the observed results.

Ever since men began to make measurements they have focused their attention on the quantity under measurement. There is always a second quantity which requires measurement. This

---

* Standard deviation of the _average_.

second quantity is the variation in the original measurements -- the scatter they exhibit about their average. The examples given above show emphatically that it is desirable to have a unit to measure this variation. Once this variation has been measured it is relatively easy to face the questions about the data that sooner or later confront every experimenter.

APPENDIX I

Critical Values for Chi Square
One in 20 Level. n Measurements

n	Lower	Upper
2	0.00393	3.841
3	0.103	5.991
4	0.352	7.815
5	0.711	9.488
10	3.325	16.919
15	6.571	23.685
20	10.117	30.144

This table is used to determine whether a group of n measurements shows (a) less variation than some standard value or (b) more variation than some standard value for the Standard Deviation.

Chi square is computed by taking the ratio of the sum of the squares of the deviations from the average to the square of the known standard deviation.

$$\text{Chi Square} = \frac{(x - \bar{x})^2}{(\text{known S.D.})^2}$$

One in 20 sets of measurements will yield values below the lower critical values, and one in 20 above the critical values even if the procedure really has the Standard Deviation entered in the denominator.

This table is extracted from larger tables available in most statistical texts. In these texts the entries will be found opposite (n - 1).

APPENDIX II

Number of measurements required to reveal whether a given procedure has a standard deviation that is a given fraction of an assigned value

Given fraction of assigned value	Chance that the indicated number of measurements will reveal the improvement	
	Four out of five	Nine out of ten
.70	28	38
.65	20	26
.60	15	19
.55	12	14
.50	9	12
.45	8	9
.40	7	8

The value of Chi square for the set of measurements, when they are obtained, is calculated as in Appendix I and is judged by the lower critical limit values. The numbers of measurements that are entered in the above table are sufficient to reduce the chance of claiming an improvement, when in reality there has been no improvement, to one in 20. The two columns refer to the chance of detecting the improvement, if actually there to the extent indicated in the first column.

For larger tables see "Selected Techniques of Statistical Analysis", edited by Eisenhart, Hastay and Wallis. McGraw-Hill Book Company, Inc., 1947.

APPENDIX III

Critical Values for the F Ratio
Each Group with n Measurements

n	F	
2	161	One chance
3	19	in ten of
4	9.28	attaining F
5	6.39	if both sets
10	3.18	come from
15	2.48	sources with
20	2.13	same standard
25	1.98	deviation.

Compute the squared standard deviation for each set and divide the larger result by the smaller to get the F ratio.

If circumstances dictate that one of the sources must have a smaller standard deviation than the other in order to be of interest and the squared standard deviation from this source made the denominator of the ratio, there is but one chance in 20 of attaining the F ratio shown above if both sources in fact have the same standard deviation.

APPENDIX IV

Number of measurements required in each of two groups to have a four out of five chance of detecting whether the standard deviation of one source is a given multiple of the other.

Multiple	No. of Measurements
1.5	39
2.0	15
2.5	9
3.0	7
3.5	6
4.0	5

The ratio of squared standard deviations is obtained by putting the value from the source expected to have the smaller standard deviation in the denominator. The ratio is judged by the critical values in Appendix III.

APPENDIX V

Factors to multiply the standard deviation of the average estimated from n measurements in order to set up a 95 percent confidence interval for the average.

n	t
2	12.706
3	4.303
4	3.182
5	2.776
10	2.262
15	2.145
20	2.093
25	2.064
∞	1.960

APPENDIX VI

Given that past experience with a process has provided a good estimate of the standard deviation of the output, the table shows the number of measurements necessary to detect whether the process average has shifted from a standard value.

Shift in Average	Measurements needed 4 out of 5	Measurements needed 9 out of 10
0.5 S.D.	30	42
1.0 S.D.	8	11
1.5 S.D.	4	5
2.0 S.D.	2	3
2.5 S.D.	2	2
3.0 S.D.	1	2

Compute

$$t = \frac{\text{standard value} - \text{average}}{\text{S.D.} / \sqrt{n}}$$

Accept shift if $t > 1.96$. Then there is one chance in 20 that there has been no shift in the process average.

The S.D. is taken from prior work.

SEE: Ferris, Grubbs, and Weaver: Annals of Mathematical Statistics, XVII, 178, 1946.