Chapter 14
Measurement Theory
Excerpted from Doeblin, "Measurement Systems"

GENERALIZED PERFORMANCE CHARACTERISTICS OF INSTRUMENTS

If you are trying to choose, from commercially available instruments, the one most suitable for a proposed measurement, or, alternatively, if you are engaged in the design of instruments for specific measuring tasks, then the subject of performance criteria assumes major proportions. That is, to make intelligent decisions, there must be some quantitative bases for comparing one instrument (or proposed design) with the possible alternatives. Now we propose to study in considerable detail the performance of measuring instruments and systems with regard to how well they measure the desired inputs and how thoroughly they reject the spurious inputs.

The treatment of instrument performance characteristics generally has been broken down into the subareas of static characteristics and dynamic characteristics, and this plan is followed here. The reasons for such a classification are several. First, some applications involve the measurement of quantities that are constant or vary only quite slowly. Under these conditions, it is possible to define a set of performance criteria that give a meaningful description of the quality of measurement without becoming concerned with dynamic descriptions involving differential equations. These criteria are called the static characteristics. Many other measurement problems involve rapidly varying quantities. Here the dynamic relations between the instrument input and output must be examined,generally by the use of differential equations. Performance criteria based on these dynamic relations constitute the dynamic characteristics.

Actually, static characteristics also influence the quality of measurement

under dynamic conditions, but the static characteristics generally show up as nonlinear or statistical effects in the otherwise linear differential equations giving the dynamic characteristics. These effects would make the differential equations unmanageable, and so the conventional approach is to treat the two aspects of the problem separately. Thus the differential equations of dynamic performance generally neglect the effects of dry friction, backlash, hysteresis, statistical scatter, etc., even though these effects affect the dynamic behavior. These phenomena are more conveniently studied as static characteristics, and the overall performance of an instrument is then judged by a semi-quantitative superposition of the static and dynamic characteristics. This approach is, of course, approximate but a necessary expedient.

STATIC CHARACTERISTICS

We begin our study of static performance characteristics by considering the meaning of the term "static calibration."

Meaning of Static Calibration

All the static performance characteristics are obtained by one form or another of a process called static calibration. So it is appropriate at this point to develop a clear concept of what is meant by this term.
In general, static calibration refers to a situation in which all inputs (desired, interfering. modifying) except one are kept at some constant values. Then the one input under study is varied over some range of constant values, which causes the output(s) to vary over some range of constant values. The input-output relations developed in this way comprise a static calibration valid under the stated constant conditions of all the other inputs. This procedure may be repeated, by varying in turn each input

considered to be of interest and thus developing a family of static input-output relations. Then we might hope to describe the overall instrument static behavior by some suitable form of superposition of these individual effects.

In some cases, if overall rather than individual effects were desired, the calibration procedure would specify the variation of several inputs simultaneously. Also if you examine any practical instrument critically, you will find many modifying and/or interfering inputs, each of which might have quite small effects and which would be impractical to control. Thus the statement "all other inputs are held constant" refers to an ideal situation which can be only approached, but never reached, in practice. Measurement method describes the ideal situation while measurement process describes the (imperfect) physical realization of the measurement method.

The statement that one input is varied and all others are held constant implies that all these inputs are determined (measured) independently of the output (should be relatively small in a good instrument), the measurement of these inputs usually need not be at an extremely high accuracy level. For example, suppose a pressure gage has temperature as an interfering input to the extent that a temperature change of 100°C causes a pressure error of 0.100 percent. Now, if we had measured the 100°C interfering input with a thermometer which itself had an error of 2.0 percent, the pressure error actually would have been 0.102 percent. It should be clear that the difference between an error of 0.100 and 0.102 percent is entirely negligible in most engineering situations. However, when calibrating the response of the instrument to its desired inputs, you must exercise considerable care in choosing the means of determining the numerical values of these inputs. That is, if a pressure gage is inherently capable of an accuracy of 0.1 percent, you must certainly

be able to determine its input pressure during calibration with an accuracy somewhat greater than this. In other words, it is impossible to calibrate an instrument to an accuracy greater than that of the standard with which it is compared. A rule often followed is that the calibration standard should be at least about 10 times as accurate as the instrument being calibrated. While we do not discuss standards in detail at this point, it is of utmost importance that the person performing the calibration be able to answer the question: How do I know that this standard is capable of its stated accuracy? The ability to trace the accuracy of a standard back to its ultimate source in the fundamental standards of the National Institute of Standards and Technology (Formerly pre-1989 National Bureau of Standards) is termed traceability.

In performing a calibration, the following steps are necessary:

I. **Examine the construction of the instrument, and identify and list all the possible inputs.**
2. **Decide, as best you can, which of the inputs will be significant in the application for which the instrument is to be calibrated.**
3. **Procure apparatus that will allow you to vary all significant inputs over the ranges considered necessary.**
4. **By holding some inputs constant, varying others, and recording the output(s), develop the desired static input-output relations.**

Now we are ready for a more detailed discussion of specific static characteristics. These characteristics may be classified as either general or special. General static characteristics are of interest in every instrument. Special static characteristics are of interest in only a particular instrument. We concentrate mainly on general characteristics, leaving the treatment of special characteristics to later sections of the text in which specific instruments are discussed.

Accuracy, Precision, and Bias

When we measure some physical quantity with an instrument and obtain a numerical value, usually we are concerned with how close this value may be to the "true" value. It is first necessary to understand that this so-called true value is, in general, unknown and unknowable, since perfectly exact definitions of the physical quantities to be measured are impossible. This can be illustrated by specific example, for instance, the length of a cylindrical rod. When we ask ourselves what we really mean by the length of this rod, we must consider such questions as these :

I. Are the two ends of the rod planes?
2. If they are planes, are they parallel?
3. If they are not planes, what sort of surfaces are they?
4. What about surface roughness?

We see that complex problems are introduced when we deal with a real object rather than an abstract, geometric solid. The term "true value," then, refers to a value that would be obtained if the quantity under consideration were measured by an exemplar method,l that is, a method agreed on by experts as being sufficiently accurate for the purposes to which the data ultimately will be put.

We must also be concerned about whether we are describing the characteristics of a single reading of an instrument or of a measurement process. If we speak of a single measurement, the error is the difference between the measurement and the corresponding true value, which is taken to be positive if the measurement is greater than the true value. When using an instrument, however, we are concerned with the characteristics of the measurement process associated with that

instrument. That is, we may take a single reading, but this is a sample from a statistical population generated by the measurement process. If we know the characteristics of the process, we can put bounds on the error of the single measurement, although we cannot tell what the error itself is, since this would imply that we knew the true value. Thus we are interested in being able to make statements about the accuracy (lack of error) of our readings. This can be done in terms of the concepts of precision and bias of the measurement process.

The measurement process consists of actually carrying out, as well as possible, the instructions for performing the measurement, which are the measurement method. (Since calibration is essentially a refined form of measurement, these remarks apply equally to the process of calibration.) If this process is repeated over and over under assumed identical conditions, we get a large number of readings from the instrument. Usually these readings will not all be the same, and so we note immediately that we may try to ensure identical conditions for each trial, but it is never exactly possible. The data generated in this fashion may be used to describe the measurement process so that, if it is used in the future, we may be able to attach some numerical estimates of error to its outputs.

If the output data are to give a meaningful description of the measurement process, the data must form what is called a random sequence. Another way of saying this is that the process must be in a state of statistical control. The concept of the state of statistical control is not a particularly simple one, but we try to explain its essence briefly. First we note that it is meaningless to speak of the accuracy of an instrument as an isolated device. We must always consider the instrument plus its environment and method of use, that is, the instrument plus its inputs. This aggregate constitutes the measurement process. Every instrument has an infinite number of inputs; that is, the causes that can conceivably affect the output,

if only very slightly, are limitless. Such effects as atmospheric pressure, temperature, and humidity are among the more obvious. But if we are willing to "split hairs," we can uncover a multitude of other physical causes that could affect the instrument with varying degrees of severity. In defining a calibration procedure for a specific instrument, we specify that certain inputs must be held "constant" within certain limits. These inputs, it is hoped, are the ones that contribute the largest components to the overall error of the instrument. The remaining infinite number of inputs is left uncontrolled, and it is hoped that each of these individually contributes only a very small effect and that in the aggregate their effect on the instrument output will be of a random nature. If this is indeed the case, the process is said to be in statistical control. Experimental proof that a process is in statistical control is not easy to come by; in fact, strict statistical control is unlikely of practical achievement. Thus we can only approximate this situation.

Lack of control is sometimes obvious, however, if we repeat a measurement and plot the result (output) versus the trial number. Figure 3.1a shows such a graph for the calibration of a particular instrument. In this instance, it was ascertained after some study that the instrument actually was much more sensitive to temperature than had been thought. The original calibration was carried out in a room without temperature control. Thus the room temperature varied from a low in the morning to a peak in the early afternoon and then dropped again in the late afternoon. Since the 10 trials covered a period of about one day, the trend of the curve is understandable. By performing the calibration in a temperature-controlled room, the graph of Fig. 3.1b was obtained. For the detection of

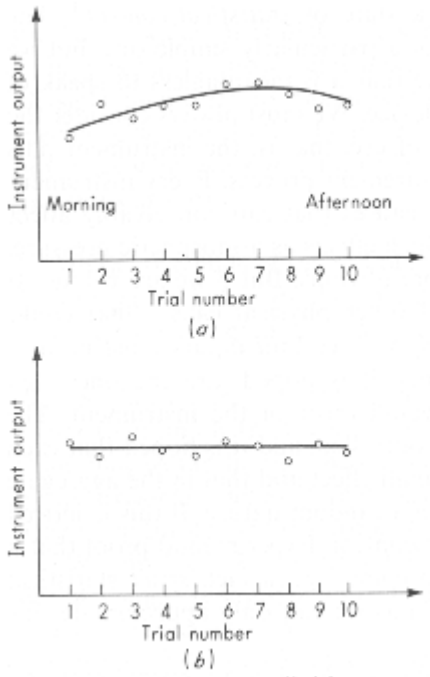

Figure 3.1 Effect of uncontrolled input on calibration.

more subtle deviations from statistical control, the methods of statistical quality-control charts are useful. If the measurement process is in reasonably good statistical control and if we repeat a given measurement (or calibration point) over and over, we will generate a set of data exhibiting random scatter. As an example, consider the pressure gage of Fig. 3.2. Suppose we wish to determine the relationship between the desired input (pressure) and the output (scale reading). Other inputs which could be significant and which might have to be controlled during the pressure calibration include temperature, acceleration, and vibration.

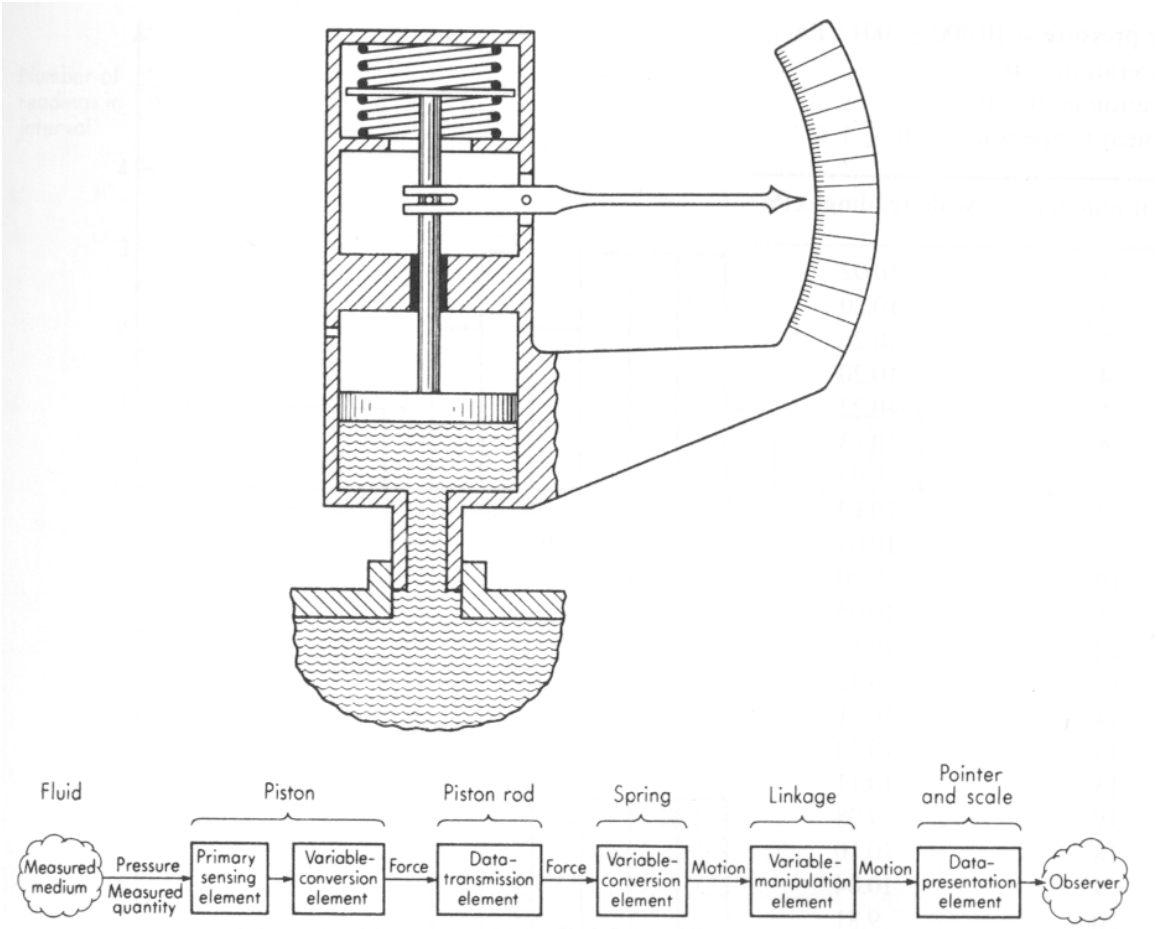

Figure 3.2 Pressure gage.

Temperature can cause expansion and contraction of instrument parts in such a way that the scale reading will change even though the pressure has remained constant. An instrument acceleration along the axis of the piston rod will cause a scale reading even though pressure again has remained unchanged. This input is significant if the pressure gage is to be used aboard a vehicle of some kind. A small amount of vibration actually may be helpful to the operation of an instrument, since vibration may reduce the effects of static friction. Thus if the pressure gage is to be attached to a reciprocating air compressor (which always has some vibration), it may be more accurate under these conditions than it would be under calibration conditions where no vibration was provided. These examples illustrate the general importance of carefully considering the

relationship between the calibration conditions and the actual application conditions.

Suppose, now, that we have procured a sufficiently accurate pressure standard and have arranged to maintain the other inputs reasonably close to the actual application conditions. Repeated calibrations at a given pressure (say, 10 kPa) might give the data of Fig. 3.3. Suppose we now order the readings from the lowest (9.81) to the highest (10.42) and see how many readings fall in each interval of, say, 0.05 kPa, starting at 9.80. The result can be represented graphically as in Fig. 3.4a.

True pressure = 10.000 ± .001 kPa
Acceleration = 0
Vibration level = 0
Ambient temperature = 20 ± 1°C

Trial number	Scale reading, kPa
1	10.02
2	10.20
3	10.26
4	10.20
5	10.22
6	10.13
7	9.97
8	10.12
9	10.09
10	9.90
11	10.05
12	10.17
13	10.42
14	10.21
15	10.23
16	10.11
17	9.98
18	10.10
19	10.04
20	9.81

Figure 3.3 Pressure-gage calibration data.

Suppose we now define the quantity Z by and we plot a "bar graph " with height Z for each interval.

$$Z \triangleq \frac{\text{(number of readings in an interval)/(total number of readings)}}{\text{width of interval}}$$

Such a "histogram" is shown in Fig. 3.4b. It should be clear from Eq. (3.1) that the area of a particular "bar" is numerically equal to the probability that a specific reading will fall in the associated interval. The area of the entire histogram must then be 1.0 (100 percent = 1.0), since there is 100 percent probability that the reading will fall somewhere between the lowest and highest values, at least based on the data available. If it were now possible to take an infinite number of readings, each with an infinite number of significant digits, we could make the chosen intervals as small as we

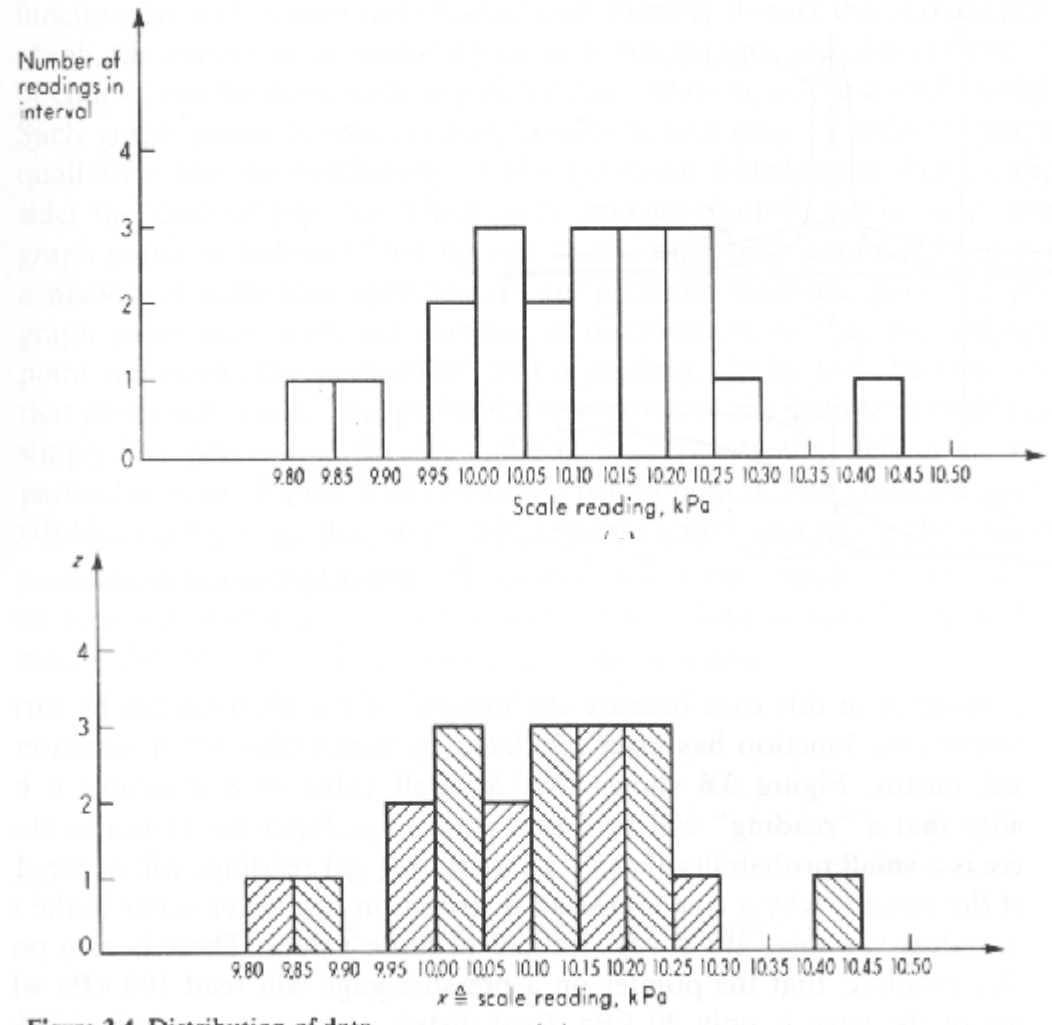

Figure 3.4 Distribution of data.

(6)

pleased and still have each interval contain a finite number of readings.

Thus the steps in the graph of Fig. 3.4b would become smaller and smaller, with the graph approaching a smooth curve in the limit. If we take this limiting abstract case as a mathematical model for the real physical situation, the function Z = f(x) is called the probability density function for the mathematical model of the real physical process (see Fig. 3.5a).

The probability information sometimes is given in terms of the cumulative distribution function F(x), which is defined by and is shown in Fig. 3.5b

$F(x) \triangleq$ probability that reading is less than any chosen value of x

$$F(x) = \int_{-\infty}^{x} f(x)\, dx$$

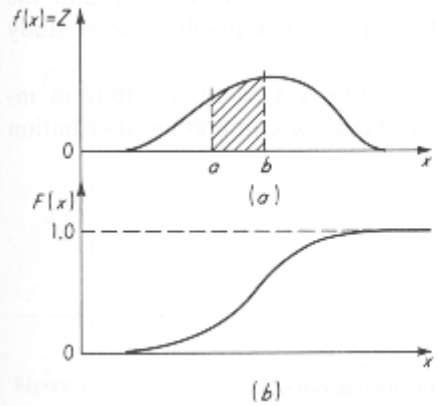

Figure 3.5 Probability distribution function.

From the infinite number of forms possible for probability density functions, a relatively small number are useful mathematical models for practical applications; in fact, one particular form is quite dominant. The most useful density function or distribution is the normal or Gaussian function, which is given by

$$f(x) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-(x-\mu)^2/(2\sigma^2)} \qquad -\infty < x < +\infty$$

Equation (3.4) defines a whole family of curves depending on the particular numerical values of μ (the mean value) and σ (the standard deviation). The shape of the curve is determined entirely by σ, with μ serving only to locate the position of the curve along the x axis. The cumulative distribution function F(x) cannot be written explicitly in this case because the integral of Eq. (3.3) cannot be carried out; however, the function has been tabulated by performing the integration by numerical means. Figure 3.6 shows that a small value of o indicates a high probability that a "reading" will be found close to μ. Equation (3.4) also shows that there is a small probability that very large (approaching +- infinity) readings will occur.

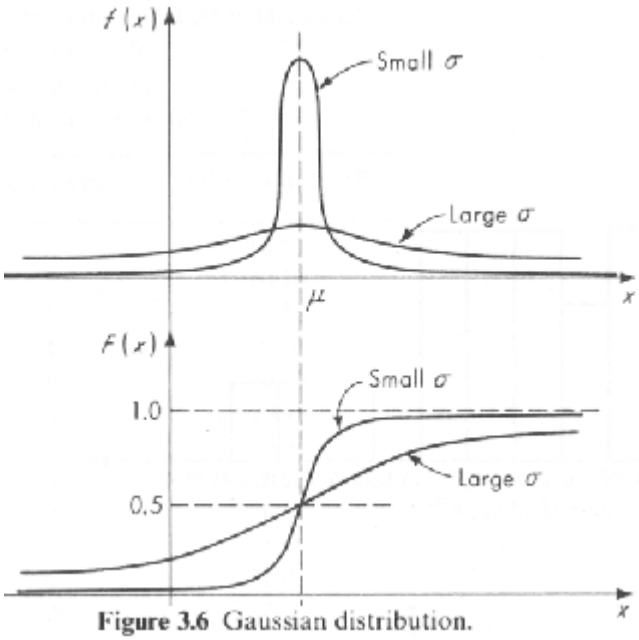

Figure 3.6 Gaussian distribution.

This is one of the reasons why a true Gaussian distribution can never occur in the real world; physical variables are always limited to finite values. There is zero probability, for example, that the pointer on a pressure gage will read 100 kPa when the range of the gage is only 20 kPa.

Real distributions must thus, in general, have their " tails " cut off, as in Fig. 3.7. Although actual data may not conform exactly to the Gaussian

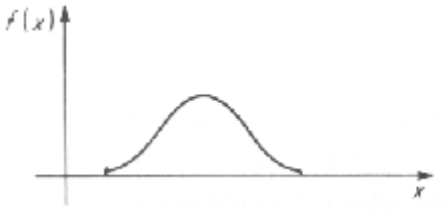

Figure 3.7 Non-Gaussian distribution.

distribution, very often they are sufficiently close to allow use of the Gaussian model in engineering work. It would be desirable to have available tests that would indicate whether the data were "reasonably" close to Gaussian, and two such procedures are explained briefly. We must admit however, that in much practical work the time and effort necessary for such tests cannot be justified, and the Gaussian model is simply assumed until troubles arise which justify a closer study of the particular situation.

The first method of testing for an approximate Gaussian distribution involves the use of probability graph paper. If we take the cumulative distributionfunction for a Gaussian distribution and suitably distort the

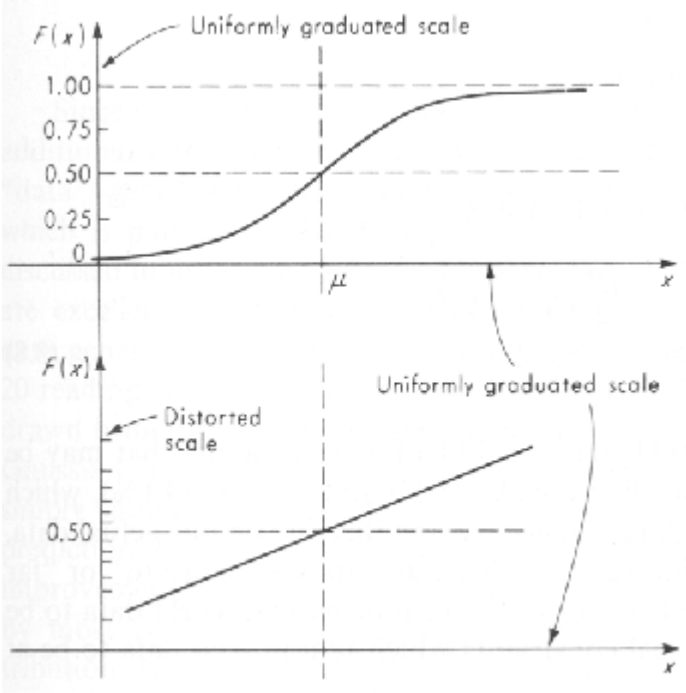

Figure 3.8 Rectification of Gaussian curve.

vertical scale of the graph, the curve can be made to plot as a straight line, as shown in Fig. 3.8. (This, of course, can be done with any curvilinear relation, not just probability curves.) Such graph paper is commercially available and may be used to give a rough, qualitative test for conformity to the Gaussian distribution. For example, consider the data of Fig. 3.3. These data may be plotted on Gaussian probability graph paper as follows: First layout on the uniformly graduated horizontal axis a numerical scale that includes all the pressure readings. Now the probability graph paper represents the cumulative distribution, so that the ordinate of any point represents the probability that a reading will be less than the abscissa of that particular point. This probability, in terms of the sample of data available, is simply the percentage (in decimal form) of points that fell at or below that particular value. Figure 3.9a shows the resulting plot. Note

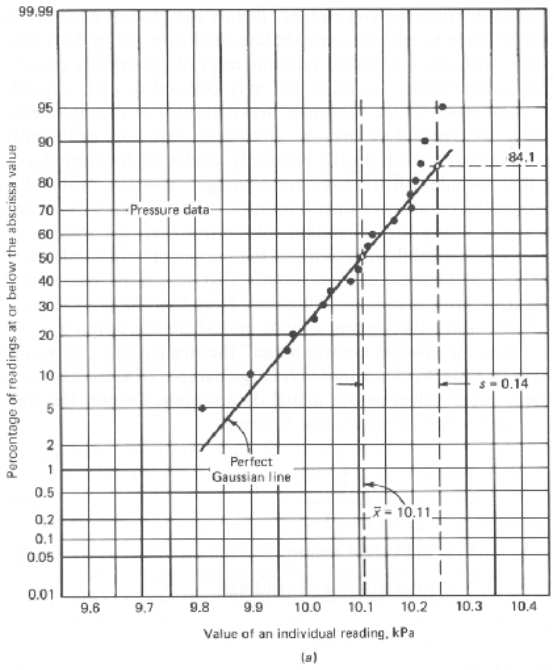

Figure 3.9 Graphical check of Gaussian distribution.

that the highest point (10.42) cannot be plotted since 100 percent cannot appear on the ordinate scale.

Also shown in Fig. 3.9a is the "perfect Gaussian line," the straight line that would be perfectly followed by data from an infinitely large sample of

Gaussian data which had the same μ and σ values as our actual data

$$\bar{X} \triangleq \frac{\sum_{i=1}^{N} X_i}{N} \tag{3.5}$$

where

$$X_i \triangleq \text{individual reading} \tag{3.6}$$

$$N \triangleq \text{total number of readings} \tag{3.7}$$

sample. To plot this line, we must estimate μ from the sample mean value X, using

and σ from the *sample standard deviation* s, using

$$s \triangleq \sqrt{\frac{\sum_{i=1}^{N} (X_i - \bar{X})^2}{N-1}}$$

The data of Fig. 3.3 give X = 10.11 and s = 0.14 kPa. Two points that may be used to plot any perfect Gaussian line are (X, 50%) and (X + s, 84.1 %), which yield the line of Fig. 3.9a for our data. Superimposing this line on our actual data, we may judge visually and qualitatively whether our data are "close to" or "far from" Gaussian. Note that there is no hope of ever proving real-world data to be Gaussian. There are always physical constraints which require real data to be at least somewhat non-Gaussian.

For a perfect Gaussian distribution, it can be shown that
68% of the readings lie within +- 1σ of μ
95% of the readings lie within +- 2σ of μ (3.9)
99.7% of the readings lie within +- 3σ of μ

Thus if we assume that our real distribution is nearly Gaussian, we might

predict, for instance, that if more readings were taken, 99.7 percent would fall within +- 0.42 kPa of 10.11. The estimates X and σ and μ are themselves random variables and can be improved by taking more readings. For example,

the standard deviation $s_{\bar{X}}$ of $\bar{X}$ may be found from[1]

$$s_{\bar{X}} = \frac{s}{\sqrt{N-1}} \tag{3.10}$$

which clearly shows a reduction in uncertainty for X as sample size N increases.

Having considered the problem of determining the normality of scattered data, we return to the main business of this section, that is, definition of the terms "accuracy," "precision," and "bias." Up to now, we have been examining the situation in which a single true value is applied repeatedly and the resulting measured values are recorded and analyzed. In an actual instrument calibration, the true value is varied, in increments, over some range, causing the measured value also to vary over a range. Very often there is no multiple repetition of a given true value. The procedure is merely to cover the desired range in both the increasing and the decreasing directions. Thus a given true value is applied, at most, twice if we choose to use the same set of true values for both increasing and decreasing readings.

As an example, suppose we wish to calibrate the pressure gage of Fig. 3.2 for the relation between the desired input (pressure) and the output (scale reading). Figure 3.13a gives the data for such a calibration over the range 0 to 10 kPa. In this instrument (as in most but not all), the input-output relation is ideally a straight line. The average calibration curve for such an instrument generally is taken as a straight line which fits the scattered data points best as defined by some chosen criterion. The most common is the least-squares criterion, which minimizes the sum of the squares of the

vertical deviations of the data points from the fitted line. (The least-squares procedure also can be used to fit curves other than straight lines to scattered data.) The equation for the straight line is taken as

$$q_o = mq_i + b \tag{3.14}$$

where

$$q_o \triangleq \text{output quantity (dependent variable)} \tag{3.15}$$
$$q_i \triangleq \text{input quantity (independent variable)} \tag{3.16}$$
$$m \triangleq \text{slope of line} \tag{3.17}$$
$$b \triangleq \text{intercept of line on vertical axis} \tag{3.18}$$

The equations for calculating m and b may be found in several references[1]:

$$m = \frac{N\Sigma q_i q_o - (\Sigma q_i)(\Sigma q_o)}{N\Sigma q_i^2 - (\Sigma q_i)^2} \tag{3.19}$$

$$b = \frac{(\Sigma q_o)(\Sigma q_i^2) - (\Sigma q_i q_o)(\Sigma q_i)}{N\Sigma q_i{}^2 - (\Sigma q_i)^2} \tag{3.20}$$

where

$$N \triangleq \text{total number of data points} \tag{3.21}$$

In this example, calculation gives $m = 1.08$ and $b = -0.85$ kPa. Since these values are derived from scattered data, it would be useful to have some idea of their possible variation. The standard deviations of m and b may be found from

$$s_m^2 = \frac{Ns_{q_o}^2}{N\Sigma q_i^2 - (\Sigma q_i)^2} \tag{3.22}$$

$$s_b^2 = \frac{s_{q_o}^2 \sum q_i^2}{N\Sigma q_i^2 - (\Sigma q_i)^2} \tag{3.23}$$

where

$$s_{q_o}^2 = \frac{1}{N} \sum (mq_i + b - q_o)^2 \tag{3.24}$$

The symbol S_{qo} represents the standard deviation of q_o. That is, if q_i were fixed and then repeated over and over, q_o would give scattered values, with the amount of scatter being indicated by S_{qo}. If we assume that this S_{qo} would be the same for any value of q_i, we can calculate S_{qo} using all the data points of Fig. 3.13a and without having to repeat anyone q_i many times. For this example; calculation gives $S_{qo} = 0.20$ kPa. Then $S_m = 0.0134$

and S_b= 0.078 kPa. Assuming a Gaussian distribution and the 99.7 percent limits (+-3s), we could give m as 1.08+-0.04 and b as -0.85+-0.24 kPa.

In using the calibration results, the situation is such that q_o (the indicated pressure) is known and we wish to make a statement about q_i (the true pressure). We should note that in computing S_{qo} either of two approaches could be used. We might use data such as in Fig. 3.13a and apply Eq. {3.24) or, alternatively, repeat a given qi many times and compute S_{qo} from Eq. {3.8). If S_{qo} is actually the same for all values of q_i {as assumed above), these two methods should give the same answer for large samples. In computing S_{qi} however, the second method is not feasible because we cannot, in general, fix q_o in a calibration and then repeat that point over and over to get scattered values of q_i .This is because q_i is truly an independent variable {subject to choice), whereas q_o is dependent {not subject to choice). Thus, in computing S_{qi} an approach such as Eq. {3,26) is necessary.

A calibration such as that of Fig. 3.13a allows decomposition of the total error of a measurement process into two parts, the bias and the imprecision {Fig. 3.13b). That is, if we get a reading of 4.32 kPa, the true value is given as 4.79 +- 0.54 kPa {3s limits), the bias would be -0.47 kPa, and the imprecision +- .54 kPa {3s limits). Of course, once the instrument has been calibrated, the bias can be removed, and the only remaining error is that due to imprecision. The bias is also called the systematic error {since it is the same for each reading and thus can be removed by calibration). The error due to imprecision is called the random error, or nonrepeatability, since it is, in general, different for every reading and we can only put bounds on it, but cannot remove it. Thus calibration is the process of removing bias and defining imprecision numerically. The total inaccuracy of the process is defined by the combination of bias and

imprecision. If the bias is known, the total inaccuracy is entirely due to imprecision and can be specified by a single number such as S_{qi}.

A more refined method of specifying uncertainty, which recognizes that S_{qi} values based on small samples {N < 30) are less reliable than those based on large samples, is available. By using the statistical t distribution, computed Sqj values are adjusted to reflect the effect of sample size. This reference gives a very comprehensive treatment of measurement uncertainty and is recommended for those wishing further details.
In actual engineering practice, the accuracy of an instrument usually is given by a single numerical value; very often it is not made clear just what the precise meaning of this number is meant to be. Often, even though a calibration, as in Fig. 3.13, has been carried out, S_{qi} is not calculated. The error is taken as the targest horizontal deviation of any data point from the fitted line. In Fig. 3.13 this occurs at $q_i = 0$ and amounts to 0.25 kPa. The inaccuracy in this case thus might be quoted as +-2.5 percent of full scale. Note that this corresponds to about +-$2S_{qi}$ in this case. This practice is no doubt due to the practical viewpoint that when a measurement is taken, all we really want is to say that it cannot be incorrect by more than some specific value; thus the "easy way out" is simply to give a single number. This would be legitimate if the bias were known to be zero {removed by calibration) and if the plus-or-minus limit given were specified as +- s, +-2s, However, if the bias is unknown (and not zero), the quotation of a single number for the total inaccuracy is somewhat unsatisfactory, although it may be a necessary expedient.

One reason for this is that if we are trying to estimate the overall accuracy of a measurement system made up of a number of components, each of which has a known inaccuracy, the method of combining the individual inaccuracies is different for systematic errors (biases) than for random errors (imprecisions). Thus, if the number given for the total inaccuracy of

a given component contains both bias and imprecision in unknown proportions, the calculation of overall system inaccuracy is confused. However, in many cases there is no alternative, and by calculation from theory, past experience, and/or judgment the experimenter must arrive at the best available estimate of the total inaccuracy, or uncertainty (as it is sometimes called), to be attached to the reading. In such cases, a useful viewpoint is that we are willing to bet with certain odds (say 19 to 1) that the error falls within the given limits. Then such limits may be combined as if they were imprecisions in calculations of overall system error.

Irrespective of the precise meaning to be attached to accuracy figures provided, say, by instrument manufacturers, the form of such specifications is fairly uniform. More often than not, accuracy is quoted as a percentage figure based on the full-scale reading of the instrument. Thus if a pressure gage has a range from 0 to 10 kPa and a quoted inaccuracy of+-1.0 percent of full scale, this is to be interpreted as meaning that no error greater than +-0.1 kPa can be expected for any reading that might be taken on this gage, provided it is "properly" used. The manufacturer mayor may not be explicit about the conditions required for "proper use." Note that for an actual reading of 1 kPa, a 0.1-kPa error is 10 percent of the reading.

Another method sometimes utilized gives the error as a percentage of the particular reading with a qualifying statement to apply to the low end of the scale. For example, a spring scale might be described as having an naccuracy of +-0.5 percent of reading or +-0.1 N, whichever is greater. Thus for readings less than 20 N, the error is constant at +-0.1 N, while for larger readings the error is proportional to the reading.

Chapter 14 continued

Combination of Component Errors in Overall System-Accuracy Calculations

A measurement system is often made up of a chain of components, each of which is subject to individual inaccuracy. If the individual inaccuracies are known, how is the overall inaccuracy computed? A similar problem occurs in experiments that use the results (measurements) from several different instruments to compute some quantity. If the inaccuracy of each instrument is known, how is the inaccuracy of the computed result estimated? Or, inversely, if there must be a certain accuracy in a computed result, what errors are allowable in the individual instruments?

To answer the above questions, consider the problem of computing a quantity N, where N is a known function of the n independent variables u_1, u_2, u_3 , ..., u_n. That is,

$$N = f(u_1, u_2, u_3, \ldots, u_n) \tag{3.28}$$

The u's are the measured quantities (instrument or component outputs) and are in error by $\pm\Delta u_1$, $\pm\Delta u_2$, $\pm\Delta u_3$, ..., $\pm\Delta u_n$, respectively. These errors will cause an error ΔN in the computed result N. The Δu's may be considered as absolute limits on the errors, as statistical bounds such as e_p's or $3s$ limits, or as uncertainties on which we are willing to give certain odds as including the actual error. However, the method of computing ΔN and the interpretation of its meaning are different for the first case as compared with the second and third. If the Δu's are considered as absolute limits on the individual errors and we wish to calculate similar absolute limits on the error in N, we could calculate

$$N \pm \Delta N = f(u_1 \pm \Delta u_1, u_2 \pm \Delta u_2, u_3 \pm \Delta u_3, \ldots, u_n \pm \Delta u_n) \tag{3.29}$$

By subtracting N in Eq. (3.28) from $N \pm \Delta N$ in Eq. (3.29) we finally obtain $\pm\Delta N$. This procedure is needlessly time-consuming, however, and an approximate solution valid for engineering purposes may be obtained by application of the Taylor series. Expanding the function f in a Taylor series, we get

$$f(u_1 \pm \Delta u_1, u_2 \pm \Delta u_2, \ldots, u_n \pm \Delta u_n) = f(u_1, u_2, \ldots, u_n) + \Delta u_1 \frac{\partial f}{\partial u_1} + \Delta u_2 \frac{\partial f}{\partial u_2} + \cdots + \Delta u_n \frac{\partial f}{\partial u_n} + \frac{1}{2}\left[(\Delta u_1)^2 \frac{\partial^2 f}{\partial u_1^2} + \cdots\right] + \cdots \tag{3.30}$$

-

$$f(u_1 + \Delta u_1, u_2 + \Delta u_2, \ldots, u_n + \Delta u_n) = f(u_1, u_2, \ldots, u_n) + \Delta u_1 \frac{\partial f}{\partial u_1} + \Delta u_2 \frac{\partial f}{\partial u_2} + \cdots + \Delta u_n \frac{\partial f}{\partial u_n} \quad (3.31)$$

So the absolute error E_a is given by

$$E_a = \Delta N = \left|\Delta u_1 \frac{\partial f}{\partial u_1}\right| + \left|\Delta u_2 \frac{\partial f}{\partial u_2}\right| + \cdots + \left|\Delta u_n \frac{\partial f}{\partial u_n}\right| \quad (3.32)$$

The absolute-value signs are used because some of the partial derivatives might be negative, and for a positive Δu such a term would reduce the total error. Since an error Δu is, in general, just as likely to be positive as negative, to estimate the maximum possible error, the absolute-value signs must be used as in Eq. (3.32). The form of Eq. (3.32) is very useful since it shows which variables (u's) exert the strongest influence on the accuracy of the overall result. That is, if, say, $\delta f / \delta u_3$ is a large number compared with the other partial derivatives, then a small Δu_3 can have a large effect on the total E_a. If the relative or percentage error Er is desired, clearly it is given by

$$E_r = \frac{\Delta N}{N} \times 100 = \frac{100 E_a}{N} \quad (3.33)$$

So the computed result may be expressed as either $N \pm E_a$ or $N \pm E_r$ %, and the interpretation is that we are certain this error will not be exceeded since this is how the Δu's were defined.

In carrying out the above computations, questions of significant figures and rounding will occur. While hand calculators allow us to easily carry many digits (without the need to think about how many are really meaningful), even here, rounding may not be entirely foolish. The tradeoff involved is between the time it takes to properly round and the time it takes (plus the greater probability of misentering a digit) to enter a long string of digits. Each individual will have to personally resolve this tradeoff. Be sure to note, however, that irrespective of what is done at intermediate steps, the final result must always be rounded to a number of digits consistent with the accuracy of the basic data.

When the delta u's are considered not as absolute limits of error, but rather as statistical bounds such as +- 3s limits, probable errors, or uncertainties, the formulas for computing overall errors must be modified. It can be shown that the proper method of combining such errors is according to the root-sum square (rss) formula

$$E_{a_{rss}} = \sqrt{\left(\Delta u_1 \frac{\partial f}{\partial u_1}\right)^2 + \left(\Delta u_2 \frac{\partial f}{\partial u_2}\right)^2 + \cdots + \left(\Delta u_n \frac{\partial f}{\partial u_n}\right)^2} \quad (3.36)$$

The overall error E_{arss}, then has the same meaning as the individual errors. That is, if Δu_i represents a +-3s limit on u_i, then Earss represents a +-3s limit on N, and 99.7 percent of the values of N can be expected to fall within these limits (if Gaussian). Equation (3.36) always gives a smaller value of error than does E_a. (3.32).

Static Sensitivity

When an input-output calibration such as that of Fig. 3.J3 has been performed, the static sensitivity of the instrument can be defined as the slope of the calibration curve. If the curve is not nominally a straight line, the sensitivity will vary with the input value, as shown in Fig. 3.16b. To get a meaningful definition of sensitivity, the output quantity must be taken as the actual physical output, not the meaning attached to the scale numbers. That is, in Fig. 3.13 the output quantity is plotted as kilopascals; however, the actual physical output is an angular rotation of the pointer. Thus to define sensitivity properly, we must know the angular spacing of the kilopascal marks on the scale of the pressure gage. Suppose this is 5 angular degrees/kPa. Since we already calculated the slope in kilopascals per kilopascal as 1.08 in Fig. 3.13, we get the instrument static sensitivity as (5)(1.08) = 5.40 angular degrees/kPa. In this form the sensitivity allows comparison of this pressure gage with others as regards its ability to detect pressure changes.

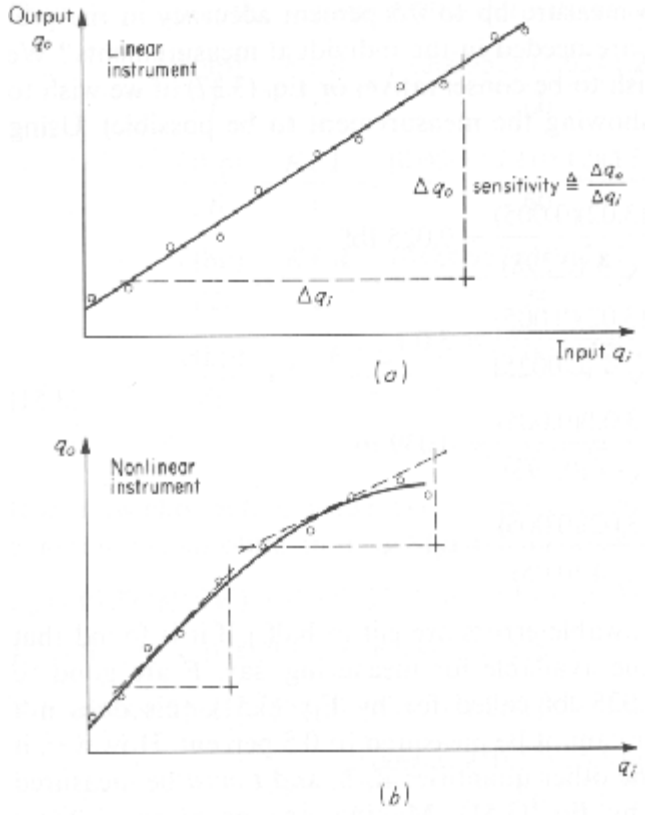

Figure 3.16 Definition of sensitivity.

While the instrument's sensitivity to its desired input is of primary concern, its sensitivity to interfering and/or modifying inputs also may be of interest. As an example, consider temperature as an input to the pressure gage mentioned above. Temperature can cause a relative expansion and contraction that will result in a change in output reading even though the pressure has not changed. In this sense, it is an interfering input. Also, temperature can alter the modulus of elasticity of the pressure-gage spring, thereby affecting the pressure sensitivity. In this sense, it is a modifying input. The first effect is often called a zero drift while the second is a sensitivity drift or scale1actor drift. These effects can be evaluated numerically by running suitable calibration tests. To evaluate zero drift, the pressure is held at zero while the temperature is varied over a range and the output reading recorded. For reasonably small temperature ranges, the effect is often nearly linear then we can quote the zero drift as, say, 0.01 angular degree/Co. Sensitivity drift may be found by fixing the temperature and running a pressure calibration to determine pressure sensitivity. Repeating this for various temperatures should show the effect of temperature on pressure sensitivity. Again, if this is nearly linear, we can specify sensitivity drift as, say, 0.0005 (angular degree/kPa)/C degree.

Figure 3,17 shows how the superposition of these two effects determines the total error due to temperature. If the instrument is used for measurement only and the temperature is known, numerical knowledge of zero drift and sensitivity drift allows correction of the readings. If such corrections are not feasible, then knowledge of the drifts is used mainly to estimate overall system errors due to temperature.

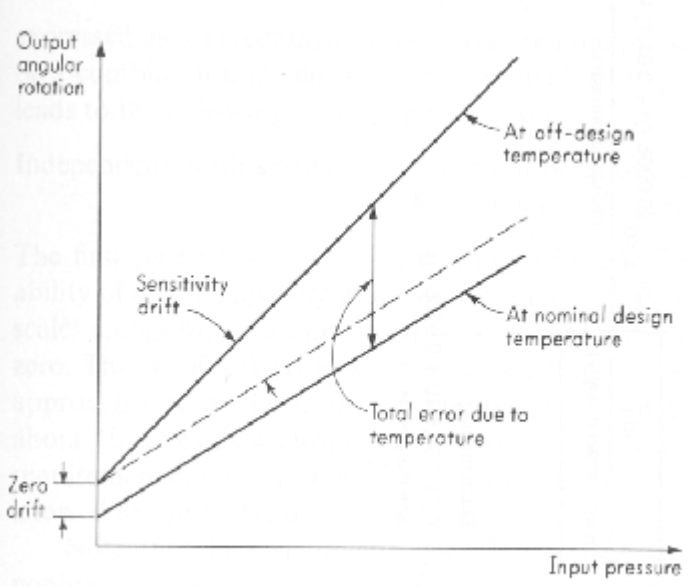

Figure 3.17 Zero and sensitivity drift.

Linearity

If an instrument's calibration curve for desired input is not a straight line, the instrument may still be highly accurate. In many applications, however, linear behavior is most desirable. The conversion from a scale reading to the corresponding measured value of input quantity is most convenient if we merely have to multiply by a fixed constant rather than consult a nonlinear calibration curve or compute from a nonlinear calibration equation. Also, when the instrument is part of a larger data or control system, linear behavior of the parts often simplifies design and analysis of the whole. Thus specifications relating to the degree of conformity to straight-line behavior are common.

Several definitions of linearity are possible. However, independent linearity seems to be preferable in many cases. Here the reference straight line is the least-squares fit, as in Fig. 3.13. Thus the linearity is simply a measure of the maximum deviation of any calibration points from this straight line. This may be expressed as a percentage of the actual reading, a percentage of full-scale reading, or a combination of the two. The last method is probably the most realistic and leads to the following type of specification:

Independent nonlinearity = +- A percent of reading or
+- B percent of full scale, whichever is greater (3.52)

The first part (+-A percent of reading) of the specification recognizes the desirability of a constant-percentage nonlinearity, while the second (+-B percent of full scale) recognizes the impossibility of testing for extremely small deviations near zero. That is, if a fixed percentage of reading is specified, the absolute deviations approach zero as the readings approach zero. Since the test equipment should be about 10 times as accurate as the instrument under test, this leads to impossible requirements on the test equipment. Figure 3.19 shows the type of tolerance band allowed by specifications of the form (3.52).

Note that in instruments considered essentially linear, the specification of nonlinearity is equivalent to a specification of overall inaccuracy when the common (nonstatistical) definition of inaccuracy is used. Thus in many commercial linear instruments, only a linearity specification (and not an accuracy specification) may be given. The reverse (an accuracy specification but not a linearity specification) may be true if nominally linear behavior is implied by the quotation of a fixed sensitivity figure.

In addition to overall accuracy requirements, linearity specifications often are useful in dividing the total error into its component parts. Such a division is sometimes advantageous in choosing and/or applying measuring systems for a particular application in which, perhaps, one type of error is more important

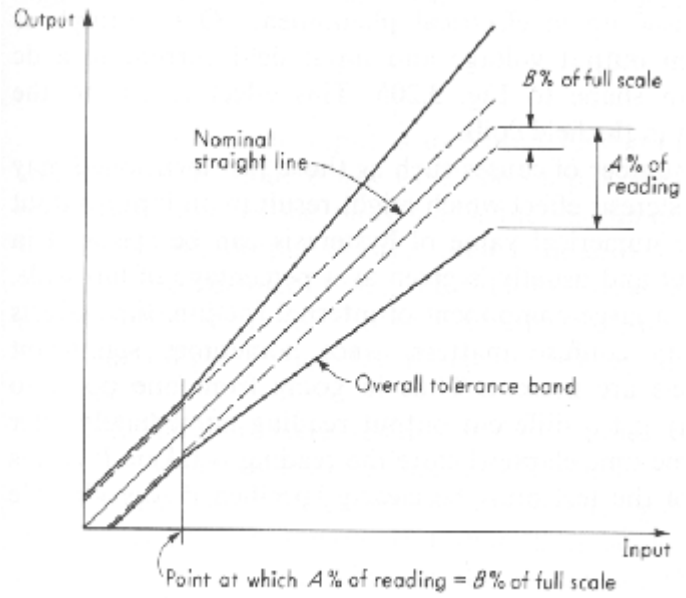

Figure 3.19 Linearity specification.

particular application in which, perhaps, one type of error is more important than another. In such cases, different definitions of linearity may be especially suitable for certain types of systems. The Scientific Apparatus Makers Association standard load-cell (force-measuring device) terminology, for instance, defines linearity as follows: "The maximum deviation of the calibration curve from a straight line drawn between no-load and full-scale load outputs, expressed as a percentage of the full-scale output and measured on increasing load only." The breakdown of total inaccuracy into its component parts is carried further in the next few sections, where hysteresis, resolution, etc., are considered.

Threshold, Resolution, Hysteresis, and Dead Space

Consider a situation in which the pressure gage of Fig. 3.2 has the input pressure slowly and smoothly varied from zero to full scale and then back to zero. If there were no friction due to sliding of moving parts, the input-output graph might appear as in Fig. 3.20a. The noncoincidence of loading and unloading curves is due to the internal friction or hysteretic damping of the stressed parts (mainly the spring). That is, not all the energy put into the stressed parts upon loading is recoverable upon unloading, because of the second law of thermodynamics, which rules out perfectly reversible processes in the real world. Certain materials exhibit a minimum of internal friction, and they should be given consideration in designing highly stressed instrument parts, provided that their other properties are suitable for the specific application. For instruments with a usable range on both sides of zero, the behavior is as shown in Fig. 3.20b. If it were possible to reduce internal friction to zero but external sliding friction were still present, the results might be as in Fig. 3.20c and d, where a constant coulomb (dry) friction force is assumed. If there is any free play or looseness in the mechanism of an

instrument, a curve of similar shape will result. Hysteresis effects also show up in electrical phenomena. One example is found in the relation between output voltage and input field current in a dc generator, which is similar in shape to Fig. 3.20b. This effect is due to the magnetic hysteresis of the iron in the field coils. In a given instrument, a number of causes such as those just mentioned may combine to give an overall hysteresis effect which might result in an input-output relation as in Fig. 3.20e. The numerical value of hysteresis can be specified in terms of either input or output and usually is given as a percentage of full scale. When the total hysteresis has a large component of internal friction, time effects during hysteresis testing may confuse matters, since sometimes significant relaxation and recovery effects are present. Thus in going from one point to another in Fig. 3.20e, we may get a different output reading immediately after changing the input than if some time elapses before the reading is taken. If this is the case, the time sequence of the test must be clearly specified if reproducible results are to be obtained.

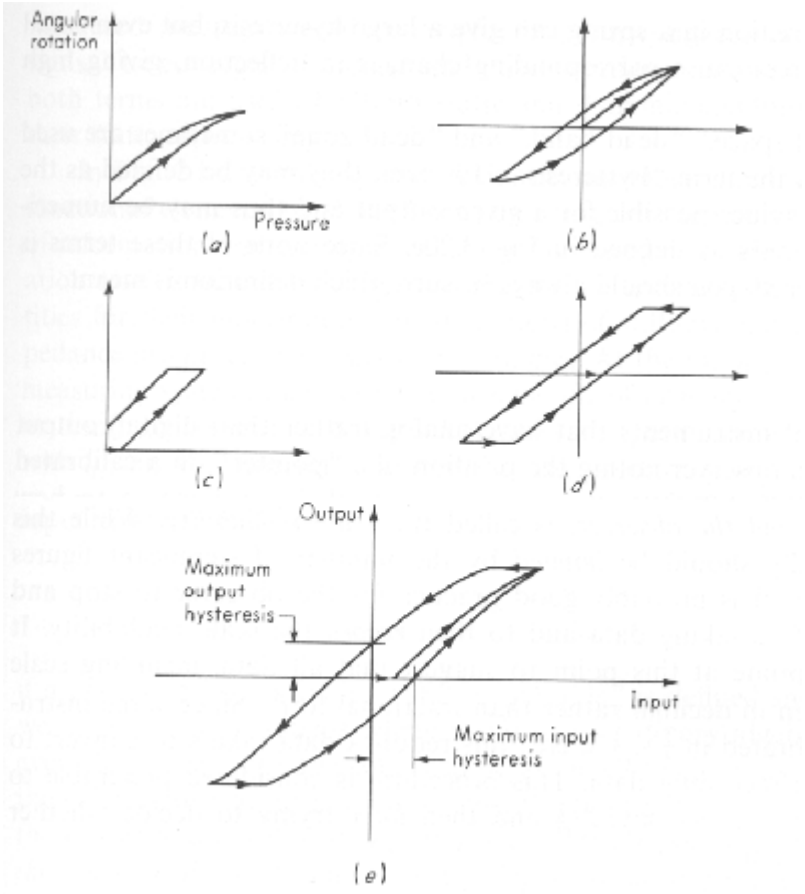

Figure 3.20 Hysteresis effects. (Magnitude exaggerated for graphical clarity.)

If the instrument input is increased very gradually from zero, there will be some minimum value below which no output change can be detected. This minimum value defines the threshold of the instrument. In specifying threshold, the first detectable output change often is described as being any "noticeable" or "measurable" change. Since these terms are somewhat vague, to improve reproducibility of threshold data it may be preferable to state a definite numerical value for output change for which the corresponding input is to be called the threshold.

If the input is increased slowly from some arbitrary (nonzero) input value, again the output does not change at all until a certain input increment is exceeded. This increment is called the resolution; again, to reduce ambiguity, it is defined as the input increment that gives some small but definite numerical change in the output. Thus resolution defines the smallest measurable input change while threshold defines the smallest measurable input. Both threshold and resolution may be given either in absolute terms or as a percentage of full-scale reading. An instrument with large hysteresis does not necessarily have poor resolution. Internal friction in a spring can give a large hysteresis, but even small changes in input (force) cause corresponding changes in deflection, giving high resolution.

The terms "dead space," "dead band," and "dead zone" sometimes are used interchangeably with the term "hysteresis." However, they may be defined as the total range of input values possible for a given output and thus may be numerically twice the hysteresis as defined in Fig. 3.20e. Since none of these terms is completely standardized, you should always be sure which definition is meant.

Scale Readability

Since the majority of instruments that have analog (rather than digital) output are read by a human observer noting the position of a "pointer" on a calibrated scale, usually it is desirable ,for data takers to state their opinions as to how closely they believe they can read this scale. This characteristic, which depends on both the instrument and the observer, is called the scale readability. While this characteristic logically should be implied by the number of significant figures recorded in the data, it is probably good practice for the observer to stop and think about this before taking data and to then record the scale readability. It may also be appropriate at this point to suggest that all data, including scale readabilities, be given in decimal rather than fractional form. Since some instrument scales are calibrated in 1/4's, 1/2's, etc., this requires data takers to convert to decimal form before recording data. This procedure is considered preferable to recording a piece of data as, say, 21 1/4 and then later trying to decide whether 21.250 or 21.3 was meant.

Span

The range of variable that an instrument is designed to measure is sometimes called the span. Equivalent terminology also in use states the "low operating limit" and "high operating limit." For essentially linear instruments, the term "linear operating range " is also common. A related term, which, however, implies dynamic fidelity also, is the dynamic range. This is the ratio of the largest to the smallest dynamic input that the instrument will faithfully measure. The number representing the dynamic range often is given in decibels, where the decibel (dB) value of a number N is defined as dB = 20 log N. Thus a dynamic range of 60 dB indicates the instrument can handle a range of input sizes of 1,000 to 1.