On the Averaging of Data

S. S. Stevens

Psycho-Acoustic Laboratory, Harvard University, Cambridge, Massachusetts

As every scientist knows, the problem of how best to average a set of data is not always easy to solve. First of all, we have a choice among the conventional measures of "central tendency" or "location," such as the mode, median, arithmetic mean, geometric mean, or harmonic mean. Each of these measures has its uses and its restrictions, but the problem is sometimes more complicated than the simple choice of one of these statistics. Complications are especially likely to arise when the quantity we want to average is a nonlinear function of the readings obtained with a particular instrument or indicator. Under these circumstances the distribution of the indicator readings is skewed relative to the distribution of the values we are interested in, and the problem becomes how to undo this skewness and the bias it produces.

Two possible procedures are available: (i) we can eliminate the skewness by making an appropriate transformation of the data, or (ii) we can bypass the problem of skewness by using a measure, such as the median or the mode, that is invariant under nonlinear monotonic transformations of the scale values. But these alternative procedures are not open to us unconditionally, for it matters considerably on what type of scale the measurements are made. Let us first examine the relationship between the various kinds of scales of measurement and the several measures of central tendency, and then let us consider an example of how we might proceed to rectify the bias in the arithmetic mean when this bias results from the use of an arbitrary measure that is nonlinearly related to the variable we are concerned with.

Scales and their statistics. The kinds of scales on which we measure things can be divided into four classes: *nominal, ordinal, interval,* and *ratio* (*1*). To each of these types of scales certain statistics are appropriate and others are not. Hence it is a matter of first importance to know which kind of scale we are dealing with. The kind of scale we work with depends, of course, upon the concrete empirical operations we are able to perform, and, as we might expect, the character of the operations determines the kind of statistics that are permissible.

This comes about because a scale erected by a given set of operations can be transformed in certain permissible ways without doing violence to the essential nature of the scale. As a matter of fact, the best way to specify the nature of a scale is in terms of its "group structure"—the group of mathematical transformations that leave the scale form invariant. And it follows quite naturally that the statistics applicable to a given scale are those that remain appropriately invariant under the transformation permitted by the scale.

This is a simple but powerful principle. It applies to all the kinds of statistics we use in the treatment of experimental data, but since we are in this article (*2*) concerned only with the statistics of central tendency we can summarize the relations between these statistics and the four kinds of scales by means of Table 1. It should be noted that the last column of the table is cumulative in the sense that a given statistic can be used with the scale indicated, as well as with all the scales listed in the rows below. The list of empirical operations in the second column is likewise cumulative, in the sense that in order to set up a given scale we need all the operations in the second column down to and including the operation listed opposite the scale.

Table 1 shows that the mode is the most general measure of central tendency. It is the only one of these statistics that may properly be used with a nominal scale, but it is a measure that may be used with all other types of scales as well. The nominal scale is the most general, or, if you like, the most

Table 1. Each of the four kinds of scales (column 1) rests on a set of empirical operations (column 2). Each scale remains invariant under certain mathematical transformations (column 3), and admits of certain measures of central tendency (column 4).

Scale	Empirical operations	Permissible transformations	Permissible measures of central tendency
Nominal	Determination of equality	Any one-to-one substitution	Mode
Ordinal	Determination of greater or less	Any increasing monotonic transform	Median
Interval	Determination of the equality of intervals or of differences	Multiplication by and addition of a constant	Arithmetic mean
Ratio	Determinations of the equality of ratios	Multiplication by a constant	Geometric mean Harmonic mean

primitive type of scale. The scale values are used only as identification tags for types or classes, such as model numbers, and with nominal scales the mode can be used to indicate which type or class has the most members in it.

The median can be used as a measure of central tendency only if the scale rests at least on an ordering operation that permits us to establish a rank order among the items measured. The classic example of an ordinal scale is the scale of hardness, where the rank order is established by the determination of what mineral will scratch what other minerals. The median may be used with ordinal scales such as hardness, and also with interval and ratio scales. As a matter of fact, the slight loss in statistical "efficiency" that results from computing the median instead of the arithmetic mean is often more than made up by the fact that the median is relatively insensitive to skewness. It is probably fair to say that the median is exploited much less than it ought to be.

Strictly speaking, the arithmetic mean is not a proper statistic for an ordinal scale, although it is often used in averaging such ordinal values as scores on tests and grades in courses. Only when we possess some operation for assuring the equality of intervals is the arithmetic mean appropriate. The classic examples of interval scales are the scales of temperature, Fahrenheit and Celsius. When we use these scales it is proper to speak of mean temperatures—as well as median or modal temperatures. The interval scale has equal units, but its zero point is arbitrary. Hence on an interval scale the computation of a geometric mean or a harmonic mean makes no sense. These two statistics are reserved exclusively for ratio scales on which the zero point is fixed. Examples of ratio scales are the everyday scales of length, weight, volume, and so forth. With ratio scales, all measures of central tendency are in principle appropriate, and which measure we decide to use must be dictated by our purpose.

The geometric mean and the harmonic mean can sometimes be used to correct the skewness in a set of data. Since these measures effect a transformation of the data, they are good for special cases that may arise in practice. The geometric mean is appropriate when the quantity we are interested in is proportional to the logarithm of the indicator reading, and the harmonic mean is appropriate when the relation is a reciprocal one.

Situations often arise, however, in which the required function is neither logarithmic nor reciprocal. The relation may in fact be such that no simple mathematical function can express it. How then do we proceed?

An experimental example. The question of how we might average indicator readings when these readings are nonlinearly related to the quantity we are trying to measure arose in an experiment on loudness. A group of 45 subjects undertook the relatively simple task of dividing a range of loudness into four equal-appearing intervals. The top end of the range was the loudness produced by a 1000-cy/sec tone about 90 db above threshold. The bottom end of the range was 40 db lower.

The subject produced the individual loudnesses by pressing one or another of five keys. Above each of the three middle keys there was a dial by means of which the subject could adjust the loudness. His task was to adjust each of the three dials until the loudness intervals between the successive tones all sounded equal to one another. He was required to test the apparent equality of the intervals by listening in both the ascending and descending order. Each subject repeated the experiment three times.

The results were recorded with the aid of a vacuum-tube voltmeter connected across the earphones. Since this was a logarithmic voltmeter, it was convenient to record the subject's settings in decibels relative to an arbitrary standard. Then with the data recorded in decibels, the question arose how to average them.

The common procedure in such instances is simply to compute the arithmetic mean of the readings as they are recorded. Second thought suggests, however, that the subjects were not listening to the logarithm of the voltage across their earphones. Rather they were listening to loudness, and their performance was in terms of the loudness heard. Since they were dividing a range into equal intervals, we can assume that the loudness they were dividing is measurable on at least an interval scale. If this is true, we are justified in taking arithmetic means of the loudness values, but, of course, we do not start with loudness values—we start with decibel readings.

It appears that the problem can be solved by an iterative process involving successive approximations. We can average the decibel values to obtain a first approximation of the relation between decibels and loudness. Then, using this approximation we can convert from decibel to loudness values and proceed to average the loudness values. This average provides a closer approximation to the relationship we seek, and using this closer approximation, we can repeat the process to determine the relationship to a still closer approximation. In this way we can make the approximation as close as we desire.

Concretely, the arithmetic means of the decibel readings were first determined and plotted (small circles in Fig. 1). The end points of the loudness range (plotted vertically) are arbitrarily called zero and 100, and the values 25, 50, and 75 mark off the equal-appearing intervals as set by the subjects. We see that over this range the subjective loudness is not a linear function of decibels. Next we draw a smooth curve (solid line) through the circles, and with the aid of this curve we change each of the original decibel readings into a corresponding loudness value. Then we average these loudness values and proceed, again via the curve in Fig. 1, to find the decibel readings that correspond to the averages of the loudness values. These new decibel readings can next be plotted in Fig. 1 to determine a new curve, the dashed curve. The process can then be repeated to determine a still better

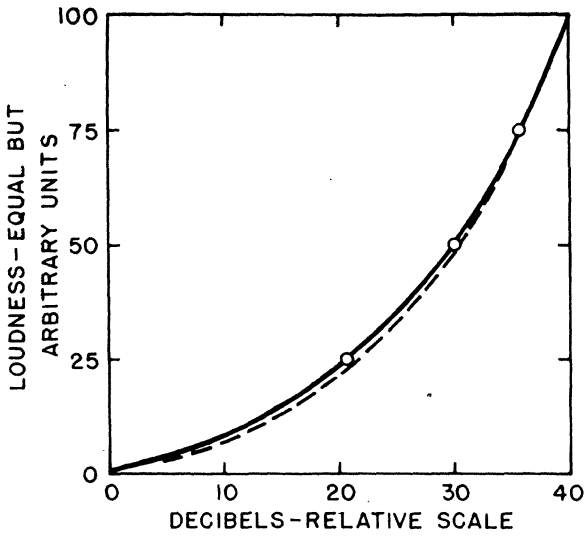

Fig. 1. Forty-five subjects adjusted three attenuators in order to divide a 40-db intensity range (abscissa) into four equal-appearing loudness intervals (ordinate). The circles, showing the arithmetic means of the decibel settings, provide a first approximation to the relation between loudness and decibels. The dashed curve shows the better approximation obtained by averaging the loudness values that were determined from the solid curve.

curve than the dashed curve, but in this particular example the new curve would be almost indistinguishable from the dashed curve.

The successive functions obtained by this procedure have greater and greater curvature, relative to that of the function determined by the original averaging of the decibel readings. The degree of the change in curvature depends on the variability of the data. If variability were nonexistent, the curvature of the function would not be changed by this process.

What we are assuming here is that it is more sensible to average loudness than to average decibels. As a further test of the reasonableness of this assumption, let us examine the forms of the distributions of the subjects' settings. Since the subject's task was to divide a fixed interval into four equal sections, it is reasonable to assume that, when measured in the proper units, the errors would distribute fairly normally, and that the greatest variability would attach to the setting that divides the total range in half—the midpoint on the ordinate of Fig. 1. We can also expect that the variability of the settings for the points 25 and 75 would be about the same size.

When we average the decibel readings, these expectations are not borne out. As shown in the upper part of Fig. 2, the histogram is broadest for the 25 point and narrowest for the 75 point. On the other hand, when we average the loudness values the distributions turn out as expected. As is shown by the bottom row of histograms in Fig. 2, the variability is greatest for the settings made to the midpoint of the loudness interval.

The experiment described here has been repeated at several intensity levels and similar results have been obtained. It has been possible to pool the data for 49 subjects who made a total of 405 settings at each of the three quarter-section points. The loudness values were averaged and the standard deviation computed and expressed as a percentage of the range of loudness the subject was working with. These mean percentage variabilities are shown in Table 2. Corresponding percentage variabilities for the decibel values themselves are also shown in Table 2. These latter values are based on only about half the data, but they are quite representative of the whole array.

In contrast to the "reasonable" behavior of the variabilities computed from the loudness values, the variabilities for the decibel readings show a drastic change along the scale. As a matter of fact, the averaging of decibels would suggest that the subjects are three times more variable when they adjust the tone to the lower quarter point than when they adjust the tone to the upper quarter point. That this unreasonable outcome can be rectified by the simple procedure of dealing with the loudness values themselves, by means of an empirical conversion function, is testimony to the importance of computing our statistics on the proper values and not on a set of arbitrary measures that are nonlinearly related to the values we are concerned with.

One further point deserves mention. In the foregoing example we have treated the data by graphical methods throughout. When, as sometimes happens, it is possible to approximate the relation between loudness and sound intensity by an analytic expression, we can dispense with graphs and use formulas. Since, in our example, loudness is approximately proportional to the cube root of the sound intensity, we can

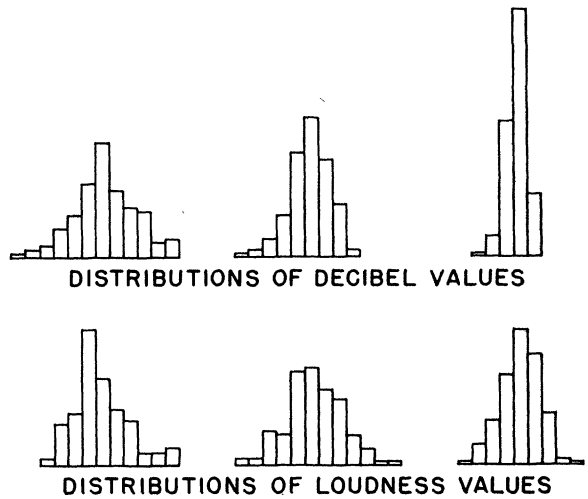

Fig. 2. Each histogram represents 135 adjustments (3 by each of 45 subjects). When measured in loudness values, the adjustments to the midpoint show the largest standard deviation (middle histogram). When measured in decibels, the lowest quarterpoint shows the largest standard deviation. The width of each bar represents 2 db or 5 loudness units, as the case might be.

Table 2. Mean percentage variabilities.

Point on loudness scale	Computed from loudness values	Computed from decibel values
25	8.3	11.8
50	9.9	8.5
75	7.9	3.9

approximate the average of the loudness values by the following procedure.

1) Divide each decibel value by 3.
2) Find the antidecibel values.
3) Average these values.
4) Find the decibel value corresponding to this average.
5) Multiply this decibel value by 3.

This procedure will undo the skewness caused by a cube-root relation between loudness and intensity. For other mathematical relations an analogous procedure can be applied.

Of course, since there are other causes of skewness than the one that concerns us here, the foregoing procedure is no panacea. For some kinds of experimental data the arithmetic mean is a poor measure of central tendency, not because the measurements are made on other than an interval or a ratio scale, but simply because out-sized errors sometimes occur in one direction or another. The resulting skewness can usually best be coped with by resort to medians (*3*).

References and Notes

1. For a fuller discussion of these scales, see S. S. Stevens, "On the theory of scales of measurement," *Science* **103**, 677 (1946); S. S. Stevens, "Mathematics, measurement and psychophysics," in S. S. Stevens, Ed., *Handbook of Experimental Psychology* (Wiley, New York, 1951), chap. 1. For other discussions of these scales, see F. B. Silsbee, "Measure for measure: some problems and paradoxes of precision," *J. Wash. Acad. Sci.* **41**, 213 (1951); C. H. Coombs, H. Raiffa and R. M. Thrall, "Some views on mathematical models and measurement theory," *Psychol. Rev.* **61**, 132 (1954); and J. P. Guilford, *Psychometric Methods*, (McGraw-Hill, New York, ed. 2, 1954).
2. Prepared under contract N5ori-76 between Harvard University and the Office of Naval Research, U.S. Navy (project NR142-201, report PNR-155).
3. See, for example, E. C. Poulton and S. S. Stevens, "On the halving and doubling of the loudness of white noise," *J. Acoust. Soc. Amer.*, in press.

Osmotic Pressure

Joel H. Hildebrand

Department of Chemistry and Chemical Engineering, University of California, Berkeley

OSMOTIC pressure no longer occupies the central role in the theory of solutions that it did a half-century ago, but in biology it retains, nevertheless, its importance as a concept, by reason of the membranes existing in living organisms. It has become evident from questions put to me that the theories of solution upon which many, if not most, biologists were brought up are now so old-fashioned that I might perform a service to the large biological clientele of this journal by presenting the subject in modern terms, emphasizing primarily the concepts involved.

The first important step in developing a theory of solutions was that made by van't Hoff, 1887, who derived the relationship that the osmotic pressure of a substance in sufficiently dilute solution is equal in magnitude to the pressure it would have if it existed as a gas in the volume occupied by the solution. This made it possible to determine the molecular weight of a nonvolatile solute by measuring its osmotic pressure in a solution of known concentration. Using this "van't Hoff law" in Carnot cycles, he derived equations relating molecular weight to the lowering of the freezing temperature and the rise in boiling temperature of the solvent. These relationships became the heart of the physical chemistry of a half-century ago. Nernst discussed electrode potentials in terms of balance between emf, solution pressure of an electrode and osmotic pressure of its ion. Every textbook of physical chemistry expounded osmotic pressure at some length, and some investigators made great efforts to measure it with precision.

But this quasi-gas model of solutions, like the first Wright airplane, with the rudder in front, proved to be a poor basis for further progress. It treated the solvent only as providing volume for the quasi-gaseous solute. It is strictly true, as van't Hoff himself pointed out, only at infinite dilution, but many investigators overlooked the restriction and applied it at concentrations where even gases cease to follow the gas laws. One enthusiast determined the freezing temperatures of concentrated solutions of calcium chloride and ascribed all deviations from the formula to the removal of part of the water from its role as solvent. The water of hydration thus calculated exceeded all the water in the vessel. He was, of course, deeply humiliated when a critic pointed this out.

One of van't Hoff's explicitly stated assumptions in deriving the equation for osmotic pressure was that the solute in dilute solution is described by Henry's law, namely, that its partial pressure is proportional to its molar concentration. This had been abundantly verified for dilute solutions of gases, and it is, indeed, almost a logical necessity; the effects of dissolved molecules too far apart to affect one another must be proportional to their number. What was not appre-