EDITORIALS

On Central Tendency and the Meaning of Mean for pH Values*

A SINGLE EXPRESSION is often used to summarize a set of data, so that a reader can get an idea of the "location" of the numbers without having to inspect the entire collection. An *index of central tendency* is the name given to the expression that serves this purpose; and three different statistical indexes are available: the mode, the median, and the mean.

Each of these indexes is determined in a different way. The *mode* is found by counting, since it represents the single item that occurs most frequently in the data. The *median* is found by ranking and counting, since it represents the middle value in the array formed when the individual items of data are arranged in ascending magnitude. The *mean* is the only summary index that requires arithmetic. It is calculated as an "average" of all of the items in the data.

Almost all investigators are aware of the distinctions just cited, but not everyone knows that an average can be calculated in at least four different ways. The decision about which way is best is the source of the current dispute about how to express the mean for a set of pH values.

The most common method of calculating a mean is used so often that most people regard it as the only kind of "mean." What it yields, however, is properly called an *arithmetic mean*: the sum of the items in the data, divided by the number of items. Thus, if a set of data contains the four values **31, 32, 33,** and **34,** the arithmetic mean is $130 \div 4 = 32.50$. In mathematical symbols, the formula for the arithmetic mean is $\Sigma x_i/n$, where x_i is any one of the items, n is the number of items, and Σ represents the act of addition.

The second method, which produces a *harmonic mean*, is used so rarely that I cannot find a good medical example of its application. It is calculated by adding the *reciprocals* of the original values, getting the arithmetic mean of the reciprocals, and converting back to its reciprocal. In the cited set of data, the four values would have a harmonic mean of $1 \div \{[(1/31) + (1/32) + (1/33) + (1/34)] \div 4\} = 1 \div \{[.03236 + .03125 + .03030 + .02941] \div 4\} = 1 \div \{.12322 \div 4\} = 32.46$. In mathematical symbols, the formula is $n/\Sigma[1/x_i]$.

The third method, which is so esoteric that I discovered its existence only while preparing this editorial, produces what is called the *quadratic mean* or the *root mean square.* It is the square root of the arithmetic mean of the squared values. The mathematical expression is $\sqrt{\Sigma x_i^2/n}$; and for our illustrative data, the quadratic mean is $\sqrt{(31^2 + 32^2 + 33^2 + 34^2)/4} = \sqrt{4230/4} = 32.52$.

The fourth method, which is uncommon but not rare, produces the *geometric mean.* Here, we multiply the n items of data together and take their nth root. For the four cited values, the geometric mean is $(31 \times 32 \times 33 \times 34)^{1/4} = \sqrt[4]{1113024} = 32.48$. Using Π as the mathematical symbol for multiplication (analogous to the Σ used for addition), the formula is $(\Pi x_i)^{1/n}$. In the days before hand calculators were readily available to do things like taking nth roots, the simplest way to get the geometric mean was to work with the logarithm of each value, find the arithmetic mean of the sum of logarithm values, and take its antilog. Thus, with our illustrative data, we would first calculate the following: $(\log 31 + \log 32 + \log 33 + \log 34)/4 = (1.49136 + 1.50515 + 1.51851 + 1.53148)/4 = 1.51163$. The antilog of the arithmetic mean of the logs yields exactly the same value produced by the multiplicative formula for the geometric mean. Thus, $10^{1.51163} = 32.48$.

All of these different ways of calculating a mean

* *The uncertainty surrounding the most appropriate method for calculation of mean pH values is exemplified by the Letter to the Editor from Drummond on page 63 of the present issue and by previous letters on the same subject from Krause (57:143-144, 1978) and from Giesecke et al (57:379-381)*—Editor.

may be interesting, but none of the mathematical strategies contains an answer to the primary question: what is the best way to represent the central tendency of a set of data? This question, alas, cannot be answered with any mathematical proof, since the answer depends on the intuition of the person examining the data. Although this intuition occurs differently to different people, most scientific workers would probably choose the median, rather than a calculated value, as the best index of central tendency. As the middle item in the ranked array of data, the median is the central point: half the items lie on one side of the median, and half on the other.

The use of a median is often mathematically unappealing, however. Any one of the four kinds of means can be easily obtained by adding, multiplying, dividing, or doing other computations with a calculator, but (unless a special program is employed) the median requires the manual work of ranking and counting. For data expressed in measured dimensions, the median is found only after the individual items are laboriously arranged in rank order, and after the number of items is counted to note the middle one. Furthermore, if n is an even number, the median falls midway between the two middle items and is arbitrarily created by splitting the difference between them. (For example, in the cited set of data, the median lies between 32 and 33 and is designated as 32.5.) Finally, the median does not lend itself to mathematical manipulation in the elegant formulas used for various statistical tests of significance.

For all these reasons, despite the median's desirability, it has not become a popular index of central tendency. Since most investigators want an index that can be calculated, what we would prefer for central tendency is a calculated mean that comes close to being a median.

In the set of four values that have been used as illustrative data, any one of the four calculated means (32.50, 32.46, 32.52, and 32.48) is satisfactory, since each one is quite close to the median value of 32.5. Problems arise, however, when the individual values in a set of data extend through a large range. Suppose our four items were **3, 30, 300,** and **3000.** What is the best way to summarize the central tendency of this set of data? Simply looking at the data, without any other mathematical considerations, we would probably say that the best index of their central tendency is a value somewhere between the two middle terms, 30 and 300. Splitting the difference would give us 165, which is the median value.

When these four items are subjected to the calculations that yield the four kinds of means, the following results emerge:

Arithmetic mean = 833.25
Harmonic mean = 10.80
Quadratic mean = 1507.56
Geometric mean = 94.87

Of these four calculated means, the geometric mean is the only one that falls in the acceptable zone (between 30 and 300) and that comes even close to the median of 165.

This feature of the geometric mean—its likelihood of providing a central tendency that approximates the median—makes it particularly attractive as a way of summarizing data that span a very wide range. The geometric mean is thus the best way of citing the central tendency for such data as bacterial counts, antibody titers, or other information in which the basic measurements get their wide range by being expressed in powers of 10 (such as 5.1×10^2, 6.8×10^3, etc.) or in powers of some other number.

With this background, we can now turn to the problem of pH. The expressions for pH are derived from a power-of-ten measurement. As the logarithm of the reciprocal of the hydrogen ion concentration, pH is $1/\log[H^+]$ or $-\log[H^+]$. To go back and forth from pH to hydrogen concentrations, we can use the formula $[H^+] = 10^{-pH}$. Thus, if the pH is 7.45, the hydrogen ion concentration is $10^{-7.45} = 3.55 \times 10^{-8}$. If $[H^+]$ is 3.16×10^{-5}, pH is $(-5) + \log 3.16 = -5 + .5 = 4.5$.

Now suppose we have a set of pH data containing the values of **1.5, 3.6, 6.7,** and **8.9.** The respective $[H^+]$ values will be $.3162 \times 10^{-1}$, $.2512 \times 10^{-3}$, $.1995 \times 10^{-6}$, and $.1259 \times 10^{-8}$. This large range of variations in $[H^+]$ will not have its central tendency well expressed with the arithmetic mean. Adding the H^+ concentrations will lead to an arithmetic mean of $.0797 \times 10^{-1}$, which yields a pH of 2.10—a poor representation of central tendency in the cited data. On the other hand, suppose we take the geometric mean of the four hydrogen ion concentrations. This mean is $[(.3162 \times 10^{-1}) \times (.2512 \times 10^{-3}) \times (.1995 \times 10^{-6}) \times (.1259 \times 10^{-8})]^{1/4} = [.001995 \times 10^{-18}]^{1/4} = [.1995 \times 10^{-20}]^{1/4} = .6683 \times 10^{-5}$. The pH of this geometric mean for $[H^+]$ is 5.18, a much better indicator of central tendency for the data, particularly since it is close to the median pH value of 5.15, which is $(3.6 + 6.7)/2$.

Now suppose we had found the mean pH simply by taking the arithmetic mean of the pH values. This mean is (1.5 + 3.6 + 6.7 + 8.9)/4 = 20.70/4 = 5.18, a result that is identical with what was obtained with the geometric mean of the $[H^+]$ values. Thus, the arithmetic mean of the pH values yields the same result that would be obtained by getting the geometric mean of the hydrogen ion concentrations, and converting it to a pH value.

Consequently, the answer to the question about how best to calculate the mean of a set of pH values is simple. Deal with the values directly, and determine their arithmetic mean. It will produce exactly the same result one would get from the optimal way, i.e. the geometric mean, of managing the data expressed in hydrogen ion concentrations.

Before concluding this commentary, I thought it might be interesting to check for observer variability among statistical consultants. I therefore telephoned Professor John W. Tukey, who had been a chemist before becoming one of the world's leading statistical authorities on data analysis. Tukey's response to the question was immediate, "Take the arithmetic mean of the pH values." His rationale, however, was considerably more ingenious than the one I have offered. Instead of referring the pH values to hydrogen ion concentrations, developing the geometric mean for the $[H^+]$ data, and converting that mean to a pH value, Tukey makes matters much more simple. Since pH is what will determine electrode potentials and other chemical potentials, Tukey suggests that pH is chemically the main entity to be considered and the hydrogen ion concentration is essentially a derivative transformation, representing the negative exponential of the pH. Thus, in Tukey's view, the "primary" measurement is the pH; and the "secondary" scale of measurement is $[H^+] = 10^{-pH}$. Since pH does not have a wide range of values, its mean can be calculated with the standard additive procedure.

Regardless of whether you prefer Tukey's rationale or mine, the recommendation is the same: add the pH's and divide by n to get their mean.

Alvan R. Feinstein, MD
Professor of Medicine and Epidemiology
Director, Robert Wood Johnson Clinical Scholar Program
Yale University School of Medicine
333 Cedar Street
New Haven, Connecticut 06510