Some Theory of Reference Values. I. Stratified (Categorized) Normal Ranges and a Method for Following an Individual’s Clinical Laboratory Values

Eugene K. Harris

The conventional population-based normal range has recently been shown to be a generally defective reference criterion for assessing individual laboratory test results. Applying a previously derived formula to published data, we find that the use of age-, sex-specific normal ranges may fail to produce a substantial improvement in sensitivity over nonspecific ranges, even when age-sex differences in mean values are statistically significant. This occurs when the difference in means is not accompanied by a sufficient reduction in the variation among individuals within a given class. Turning therefore to comparison of an individual’s current measurement with his own previous value(s), I suggest a simple statistical model that leads to sequential testing of each new observation against an exponentially weighted moving average of previous results. Estimates of biological and analytical components of variance are required. The ability of this method to detect trends in very short series is explored with the aid of computer-simulated laboratory data. A sample of these data is also used to illustrate the application of these estimation and testing procedures by means of a graph.

Additional Keyphrases: biochemical individuality ∙ statistics ∙ variation, source of ∙ components of variance

The appropriateness of the population-based normal range as a reference for interpreting an individual measurement depends on the ratio of intra-to interindividual variation in the constituent measured (1). The average value of this ratio may be defined as

\[ r = \left[ E \sigma_i^2 / \text{Var} \mu_i \right]^{1/2} \]

where \( E \sigma_i^2 \) denotes the average value of individual variances \( \sigma_i^2 \), and \( \text{Var} \mu_i \) represents the variance of mean values. A person’s own ratio is \( r_i = \sigma_i / (\text{Var} \mu_i)^{1/2} \). If \( r \leq .6 \), the conventional normal range will almost always be less sensitive than desired to changes in biochemical status. For example, in this case, a 95% normal range would include at least 99.99% of the average normal individual’s distribution of values over time. In other words, a deviation from his usual mean larger than 3.9 times his standard deviation would be required to place him outside the conventional normal range. For individuals whose mean values are not equal to the population mean, the required changes in standard deviation units would differ, of course, but in the large majority of cases such changes would exceed two standard deviations.

As the ratio \( r \) increases, particularly as \( r \) exceeds 1.4, the normal range becomes a more trustworthy reference, at least for the individual whose standard deviation is equal to the average value, \( (E \sigma_i^2)^{1/2} \). The normal range remains less sensitive than desired in persons whose intra-individual variations are smaller than average \( (r_i < r) \), but overly sensitive in those whose \( \sigma_i \) values are larger than average \( (r_i > r) \). Nevertheless, when only a single measurement of some constituent has been obtained (e.g., after initial examination at a screening center), a population-based normal range may be the only reference available. It seems worthwhile, therefore, to consider refinements to the normal range that will improve its usefulness as a guide by increasing the corresponding \( r \)-value.

Some increase in \( r \) is usually achieved by stratifying normal values by age and sex, since this tends to decrease \( \text{Var} \mu_i \) within each stratum. In the first part of this paper, the usefulness of stratified normal ranges will be assessed for some examples of published data where the statistical significance of age-related or sex-related (or both) differences has been clearly shown. In the second part of the paper, a statistical procedure will be considered for analyzing
short series of observations from the same individual, that is, use of the individual’s own past record as a reference for interpreting a current measurement.

### Age- and Sex-specific Normal Ranges

As stated in (1), if \( \sigma_1 \) and \( r_1 \) are the (single-sample) standard deviation and \( r \)-value, respectively, before age, sex (or other) classification, and if \( \sigma_2 \) and \( r_2 \) are the average within-class values after classification, then \( \frac{\sigma_2}{\sigma_1} = \left( \frac{r_1/r_2}{1 + r_1^2} \right)^{1/2} \). Setting \( r_2 \) in this formula equal to a desired value (say, 1.4 or greater), the percentage reduction in standard deviation that must then be achieved by classification may be calculated for a given value of \( r_1 \). Some results of these calculations are listed in Table 1. Values of \( r \) before classification frequently lie between .8 and 1.0, although as analytical methods become more precise the \( r \)-value for many blood constituents could fall as low as 0.6 or even lower [see (1), Table 2]. Hence a reduction in standard deviation of at least 20 to 25% (i.e., \( \left[ \frac{\sigma_1 - \sigma_2}{\sigma_1} \right] \geq .25 \)) would be hoped for as a result of age, sex classification. Published statistics indicate that such a reduction is often not attained, even in cases where the differences in mean values by age or sex, or both, are highly significant statistically. Consider the following examples.

### Cholesterol and Alkaline Phosphatase in Women, Pre- and Post-Menopausal Age Groups

Wilding et al. (2, Table IV) present the following mean and standard deviations of serum cholesterol in women, by age. These individuals attended a well-population screening center, in which the AGA Autochemist was used.

<table>
<thead>
<tr>
<th>Age group, y</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>20–29</td>
<td>2040</td>
<td>359</td>
</tr>
<tr>
<td>30–39</td>
<td>2213</td>
<td>404</td>
</tr>
<tr>
<td>40–49</td>
<td>2364</td>
<td>388</td>
</tr>
<tr>
<td>50–59</td>
<td>2630</td>
<td>474</td>
</tr>
<tr>
<td>60–69</td>
<td>2738</td>
<td>398</td>
</tr>
</tbody>
</table>

The numbers of women in each age group ranged from 72 in the youngest decade to 283 in the 40–49 year group. To keep the calculations simple, I have ignored these differences and have assumed that the reported statistics represent true values.

The average cholesterol in the women 20–49 years of age was 2206 mg/liter, increasing to 2884 in those who were 50–70 years old. With more than 500 women in each group, the statistical significance of this increase is beyond question. The average single-sample standard deviation within each of these two age groups was 413 mg/liter. However, if we now combine age classes and compute the standard deviation prior to stratification, we obtain the value 482.\(^1\) Despite the large increase in mean level from pre- to post-menopausal ages, stratification by age has reduced the single-sample standard deviation by only 14%. This percentage reduction is probably not sufficient to make an age-specific normal range for cholesterol in women appreciably more useful as a reference criterion than a nonspecific range. The nonspecific \( r \)-value reported for cholesterol in ref. 1 was extremely small, .4–.6, because of a very high degree of interindividual variation. Table 1 indicates that a percentage reduction in standard deviation as a result of age, sex classification would have to exceed 40% to warrant the use of stratified normal ranges.\(^2\)

Werner et al. (3), in their study of ambulatory and apparently healthy individuals in San Francisco, showed approximately the same mean difference in cholesterol as was shown in ref. 2 between pre- and post-menopausal ages in women, and these authors did not publish standard deviations for cholesterol and other constituents for which the distributions appeared lognormal rather than gaussian. Would our conclusions from the data of Wilding et al. have been altered if the means and standard deviations of the logarithms had been analyzed? There are simple formulas for converting the arithmetic mean and standard deviation of a lognormal distribution to the corresponding statistics of the logarithms.\(^3\) Applying these formulas to the cholesterol data of Wilding et al. again showed a 14% decrease in the standard deviation.

In the study of Wilding et al., post-menopausal women showed a 40% higher mean value of alkaline phosphatase than did women of pre-menopausal age. Because the coefficient of variation was about the same in both groups, the arithmetic mean and standard deviation are.

---

1. Given \( m \) classes with equal numbers of individuals, means \( s_k \) and variances \( s_k^2 \) (\( k = 1, 2, \ldots, m \)), the variance of the combined data may be calculated from the formula,

\[
s^2 = \left[ \frac{\sum_{k=1}^{m} s_k^2 + \sum_{k=1}^{m} (s_k - \bar{s})^2}{m} \right] / m
\]

2. The means for cholesterol in men and women of all ages were close in the data of Wilding et al., so that a sex-specific range alone would offer no improvement over a nonspecific range.

3. If \( \log x \) is gaussian-distributed with mean \( \mu \) and variance \( \sigma^2 \); then \( E x = e^{\mu+\sigma^2/2} \) and \( \text{Var } x = (e^{2\sigma^2}-1)\text{Var } x \). Thus, if \( x \) and \( s^2 \) are estimates of \( E x \) and \( \text{Var } x \), respectively, then

\[
\text{Est. } \sigma^2 = \log_e \left[ (s/2)^2 + 1 \right], \quad \text{and}
\]

\[
\text{Est. } \mu = \log_e x - \left( \text{Est. } \sigma^2 \right)/2
\]
standard variation were converted to their logarithmic counterparts, given in the table below.

<table>
<thead>
<tr>
<th>Age class, y</th>
<th>Alkaline phosphatase, in log&lt;sub&gt;10&lt;/sub&gt; King–Armstrong units</th>
</tr>
</thead>
<tbody>
<tr>
<td>20–49</td>
<td>0.849</td>
</tr>
<tr>
<td>50–69</td>
<td>0.989</td>
</tr>
</tbody>
</table>

The range (and SD) of age-class means in Leonard’s data is more than twice as great as in the data of Wilding et al., while the average SD within each age class is smaller by about a third. These conditions (but primarily the greater spread in age-class means) produce a 30% decrease in standard deviation because of age classification in the Leonard study [(0.22 – 0.15)/0.22], as compared with a 4% decrease in the study of Wilding et al. The former figure justifies (at least on statistical grounds) replacing a single non-specific normal range in serum albumin by a set of age-specific ranges; the latter result certainly does not. Yet in both sets of data, the observed slopes were highly significant statistically because of the large numbers of individuals sampled. I discuss this point in more detail below.

Sex-related Difference in Uric Acid

It is established that males, on the average, show significantly higher concentration of serum uric acid than do females. The studies of Wilding et al. (2), Werner et al. (3), and Reed et al. (4) also indicate small but statistically significant post-menopausal increases in women. Ignoring this for the moment, the following unweighted means and standard deviations may be calculated from the data of Wilding et al.:

<table>
<thead>
<tr>
<th>Uric acid, mg/liter</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td>Males (20–69 y)</td>
</tr>
<tr>
<td>Females (20–69 y)</td>
</tr>
</tbody>
</table>

The average within-sex standard deviation is 11.6. Pooling all observations yields an overall standard deviation of 13.8 mg/liter. Therefore stratification by sex has decreased the standard deviation by 16%. Further subdividing the data for females into age groups 20–49 y (mean, 4.73; SD, 1.0) and 50–69 y (mean, 5.35; SD, 1.16) produces a standard deviation that is 18% smaller than before any stratification. If we assume that the initial r-value is only slightly less than 1.0 (1, Table 2), this reduction is barely sufficient to justify specific normal ranges for men and pre- and post-menopausal women.

Effect of Age on Values for Serum Albumin

Highly significant declines in serum albumin with advancing age in both sexes have been frequently reported. The following two sets of data provide an interesting contrast.

<table>
<thead>
<tr>
<th></th>
<th>Wilding et al. (2)</th>
<th>Leonard (5)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SD</td>
</tr>
<tr>
<td>Albumin, g/liter, averaged for both sexes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;30</td>
<td>43.6</td>
<td>2.2</td>
</tr>
<tr>
<td>30–39</td>
<td>43.0</td>
<td>2.3</td>
</tr>
<tr>
<td>40–49</td>
<td>42.4</td>
<td>2.2</td>
</tr>
<tr>
<td>50–59</td>
<td>42.1</td>
<td>2.1</td>
</tr>
<tr>
<td>60–69</td>
<td>41.7</td>
<td>2.2</td>
</tr>
<tr>
<td>av</td>
<td>2.2</td>
<td></td>
</tr>
<tr>
<td>SD for all ages</td>
<td>2.3</td>
<td>2.2</td>
</tr>
</tbody>
</table>

These examples demonstrate that differences among mean values by age or sex that are large enough to attain a high degree of “statistical significance” may be too small to justify replacing a single overall normal range with several age- and sex-specific (e.g.) ranges. This should not be surprising, because any difference between means would become statistically significant if the sample sizes were large enough. Thus, the central question is: How large do such differences have to be to warrant age-, sex-specific ranges as reference values for single observations? The single sample standard deviation before stratification ($\sigma_1$) is related to the average within-stratum standard deviation ($\sigma_2$) by the general formula, $\sigma_2^2 = \sigma_1^2 + \text{Var } \mu_g$, where $\mu_g$ is the mean of individuals in the g-th stratum. Let $k_c = (\sigma_1/\sigma_2)_c$ denote the critical ratio of “before” to “after” standard deviations required to raise the pre-stratification value of $r (r_1)$ to a post-stratification value of 1.4 in each stratum. For the $r_1$-values listed in Table 1, $k_c$ may be calculated by subtracting each proportionate reduction from unity and inverting the result; $k_c$ will always exceed unity. Then, $\text{Var } \mu_g / \sigma_2^2$ must be greater than $k_c^2 - 1$. For two classes, a and b, $\text{Var } \mu_g = (\mu_a - \mu_b)^2/4$; thus, the required condition is $[\mu_a - \mu_b]^2/\sigma_2^2 \geq 4(k_c^2 - 1)$. Substituting for $\mu_a$, $\mu_b$ and $\sigma_2^2$, the observed statistics $x_a$, $x_b$ and a pooled intraclass variance $s^2$, based on sample sizes $n_a$ and $n_b$, the left-hand side of this inequality becomes the square of Student’s t-statistic, divided by $n_a n_b/(n_a + n_b)$.

For example, suppose $r_1 = .8$. Then, $k_c = [1/(1 - .23)] = 1.3$, and $t^2$ must exceed 2.76 $n_a n_b/(n_a + n_b)$. Suppose, further, that $n_a = n_b = 50$, individual in each stratum. Then $t^2$ must exceed 69, or $t$ must exceed 8.3! Such a t-value is much larger than that required for a “statistically significant” difference between two means at the 5%, 1%, or even the .1% level. The disparity would be even greater for larger sample sizes. Because linear regression is tested for significance by one degree of freedom in the analysis of variance, and the F-test with 1 and n degrees of freedom is equivalent to $t^2$ with n degrees of freedom, the same numerical criteria apply to observed
trends with age, as in the data for albumin cited above.4

To sum up, the criterion of variance reduction proposed here (Table 1) is clearly more stringent than the usual significance test. It implies that statistically significant differences in mean values are necessary, but may not be sufficient, to establish (through separate reference ranges) the diagnostic importance of age, sex or other demographic attributes. Age- and sex-related effects resting on well-founded physiological bases should be able to meet this additional test without difficulty. For example, the differences between pre-pubertal, adolescent, and adult values for serum alkaline phosphatase activity (3) easily pass.

Reference Values from an Individual's Past Record

Let us consider now the statistical problems involved in assessing a person’s current measurement by referencing his own past record rather than a population based normal range. Suppose that some blood constituent is being measured in an individual at equal intervals of time. As each new measurement is obtained, we ask: (a) Does this new measurement differ significantly (statistically) from what we might have expected, given the measurements previously recorded? (b) If not, how should we combine present and past measurements to obtain the most precise estimate of the true present value of the constituent?

The methods suggested below for answering these questions derive from a simple mathematical model that has been used extensively in industrial process control and economic forecasting.

Assume that the biological state affecting the measured quantity is fluctuating in a random manner so that the true value at time t, say , is equal to the preceding true level plus a random shift , i.e.,

\[ m_t = m_{t-1} + \Delta_t, \tag{1} \]

where, is the net result of physiological fluctuations during the interval from t - 1 to t. In steady state, and Var is the variance of shifts in the actual amount of the constituent from one sampling time to the next in the i-th individual. The measurement of , say , is subject to analytic error , so that

\[ x_t = m_t + a_t. \tag{2} \]

We assume that and denote Var by . The assumption of zero mean ignores possible non-

The mathematical properties and implications of the model specified by equations 1 and 2 have been discussed in detail by Stewart (6); also, see Muth (7). In closely controlled biochemical variables, where only narrow ranges are compatible with life, the variance will be very small relative to .

We should mention an alternative model in which equation 1 is replaced by the equation, , where is an overall mean and is a physiological fluctuation of the true value from this mean. This equation is a statement of homeostasis over a period of time, although it, too, permits fluctuations in the actual amounts present at given sampling points during the overall time period. It is this model that has been used in past studies (e.g., 8), and estimates of intrapersonal variance in those studies denoted by have referred to the variance of . Under this model, all observations would be equally weighted in estimating the mean . I believe equation 1 is more appropriate to the problem considered here, where attention is focussed on the true value at a specific time and the possibility that this value may represent a nonrandom change from past experience. Thus, the probability of the current level of a constituent conditional on previous levels is of importance here. Moreover, the earlier studies were concerned with weekly observations over a period of several months, whereas we are thinking here of a much wider range of time periods: for example, annual or semi-annual screening examinations or, on the other hand, daily measurements in a clinical situation that may be changing rapidly. In such contexts, the concept of an overall mean is not relevant. For this reason, equations 1 and 2 are sometimes referred to in

Table 2. Optimal Weighted Averages for Estimating Based on Equations 1 and 2a

<table>
<thead>
<tr>
<th>No. observations (t)</th>
<th>Weighted average formulas for weights</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>( s_1 = x_1 )</td>
</tr>
<tr>
<td>2</td>
<td>( s_1 = w_x x_1 + (1 - w_x)x_i )</td>
</tr>
<tr>
<td>2</td>
<td>( s_2 = w x_t + (1 - w)s_{t-1} )</td>
</tr>
</tbody>
</table>

\[ w = \left( -c_1 + \sqrt{(c_1 + 2) - 4} \right) / 2 \]  

(a) Source: Stewart (6)

(b) For \( t > 2 \) and \( c_1 > .25 \), is given to a sufficiently close approximation by the formula

\[ w = \left[ -c_1 + \sqrt{(c_1 + 2)^2 - 4} \right] / 2 \]

This equation has been graphed in Figure 1. For \( c_1 < .25 \), the exact formula, \( w = \left[ (w_{t-1} + c_t)/(w_{t-1} + c_t + 1) \right] \) should be used.

---

4 As pointed out by a reviewer of this paper, if X is a continuous variable (possibly grouped) like age, and Y is a laboratory measurement of a clinical variable, then, if X and Y are linearly related with correlation coefficient , the variance of Y given X (corresponding to within-stratum variance) equals \((1 - \rho^2)\) times the overall (unstratified) variance of Y. Hence, to achieve a 25% reduction in the standard deviation of Y by stratifying into a series of X-classes requires that \( \rho \) exceed 0.66. Such a correlation is not often found between a clinical variable and age or some other continuous attribute.

5 In practice, unfortunately, any measure of biological variance must also include the possibility of long-term analytical variance.
the statistical literature as representing a nonstationary, “random walk” model (e.g., 9). When physiological fluctuations are small relative to analytical fluctuations, the two models become very similar, and in this case (as we point out later), equation 1 leads to equal weighting of the observations. In general, however, Var $\Delta t$ exceeds Var $e$. The estimate of Var $\Delta t$ is based on the variance of differences between successive observations, according to the formula,

$$\text{Est } \text{Var } \Delta t = \text{Est } \sigma^2_{\Delta t} = \text{Var } d_t - 2 \sigma^2_*,$$

where $d_t = x_t - x_{t-1}$.

How does this hypothetical model (equations 1 and 2) relate to questions $a$ and $b$ posed above? The model postulates that the fluctuations observed in a time series of measurements $x_1, x_2, \ldots, x_t$ are the effects of many random biological and analytical factors that were operating before measurement began and are continuing. It is, therefore, a mathematical statement of the so-called null hypothesis under which the assumption is made that none of these fluctuations have arisen from a specific causal factor that the investigator himself may have introduced or about which he would be particularly concerned (e.g., a newly developed pathological process).

Suppose for the moment that the null hypothesis represented by this model is true for the $t$ measurements obtained to date. That is, suppose the answer to question $a$ is negative for every observation from $x_2$ to $x_t$, and we turn to question $b$. Stewart (6) has pointed out that in general the most precise estimate of the true value at time $t$ is not $x_t$ alone but a weighted average of all the observations, $x_1, x_2, \ldots, x_t$. The weights specified by the model defined in equations 1 and 2 may be summarized as shown in Table 2. The parameter $c_i$ equals $\sigma^2_{\Delta t}/\sigma^2_*$. For simplicity, the subscript $i$ has been omitted from all terms except $c_2$ and $\sigma^2_{\Delta t}$.

The form of the weighted average given in Table 2 for $t > 2$ may be recognized as the formula for computing an exponentially weighted (or geometric) moving average of the $x$’s (6, p 249). This simple recursive form permits easy computation of the successive averages $s_3, s_4, \ldots$ A good general reference to exponential smoothing is Brown’s text (10).

---

Figure 1 graphs the values of $w$ from equation 3 (Table 2) for $c_i > 0.25$. On a semi-log grid, the graph is approximately linear up to $c_i = 4$. Also included is a graph of $w_2$ whose formula is given in the last column of Table 2. As $c_i < 0$, $w \to 1/t$, weighting all observations equally. This is reasonable because in the absence of biological variation, all observations become equally good estimators of $m_t$. As $c_i \to \infty$ (i.e., when biological variance greatly exceeds analytical “noise”), $w \to 1$, putting almost all the weight on $x_t$ because only the current observation estimates $m_t$ without intervening biological fluctuation.

The observations $x_1, x_2, \ldots x_t$ taken as estimators of their respective true states $m_1, m_2, \ldots m_t$ have identical variances $\sigma^2_*$ and are statistically independent because their analytical deviations are independently distributed. However, when the weighted average $s_t$ is used to obtain a precise estimate of $m_t$, then each of the observations contained in $s_t$ is also considered an estimator of $m_t$. In this role, their variances are not identical. For example, the variance of $x_2$ as an estimator of $m_2$ is $\sigma^2_*$, but the variance of $x_1$ as an estimator of $m_2$ is $\sigma^2_1 + \sigma^2_2 = \sigma^2_*$ (1 + c), allowing for the variance of the biological fluctuation between the two measurement times. Following this reasoning, it can be shown (6) that $\text{Var } s_2 = w_2 \sigma^2_*$ and that, in general, when $s_t$ is used to estimate $m_t$, the variance of $s_t$ is $w_t \sigma^2_* = w \sigma^2_*$ for $t > 2$ when $c > 0.25$.

Estimation of $m_t$ is not the only function that $s_t$ serves. Under the null hypothesis, it also becomes a forecast of the mean value at time $(t + 1)$. When used for this purpose, $\text{Var } s_t = w_t \sigma^2_* + \sigma^2_3 = \sigma^2_3(w_t + c)$. This brings us back to question $a$: sequential testing of each new observation, say $x_t$, against its forecasted value $s_{t-1}$ under the null hypothesis. The suggested test statistic is the standard gaussian (or normal) deviate,

$$z_t = |x_t - s_{t-1}|/[\text{Var } x_t + \text{Var } s_{t-1}]^{1/2} = |x_t - s_{t-1}|/\sigma (w_{t-1} + c + 1)^{1/2}$$

(4) declared significant at the 5% level if $z_t$ exceeds 1.96.
At the beginning of observation of an individual, no prior information exists concerning the variance \( \sigma^2_{A(i)} \). At best, an average value of \( \text{Var} \ e_i = (\sigma^2_{A}) \) may be available from the literature (e.g., 8, 11). As noted above, this is likely to underestimate \( \sigma^2_{A(i)} \), and hence the ratio \( c_i \). However, use of the wrong value of \( c \) in equation 4 should not have too great an effect on the statistical test. For example, even if only half the correct value of \( c_i \) were used, the value of \( z_t \) given the same value of \( (x_t - s_{t-1})/\sigma_A \), would be increased by 10–20%. After at least four or five observations have been obtained without rejecting the null hypothesis on sequential testing (see below), the variance of successive differences of these measurements may be used to estimate \( \sigma^2_{A(i)} \), from which values of \( c_i \) and \( w \) specific to the individual may be calculated.

In the following sections I consider: (a) the sensitivity ("power") of the normal deviate test, in particular its ability to detect a linear trend within a few observations; (b) a computer-generated example to provide data for checking results derived from mathematical formulas; and (c) use of a graph to aid in applying the estimation and testing procedures described here.

### Power of the Normal Deviate Test

As each new observation is compared with the individual’s past record, the null hypothesis being tested is that \( E x_t = E s_{t-1} \), using \( s_{t-1} \) as a predictor of the true level at time \( t \). The alternative hypothesis claims that at least part of the difference between \( x_t \) and \( s_{t-1} \) is "nonrandom," due to a specific causal factor. If \( D \) represents this nonrandom change, then under the alternative hypothesis \( E x_t = E s_{t-1} + D \).

If \( D \neq 0 \), how likely is the normal deviate test to recognize this by rejecting the null hypothesis? The probability of correct rejection of the null hypothesis is called the "power" of the statistical test and is denoted by \( 1 - \beta \), where \( \beta \) is the probability of false nonrejection or "error of the second kind." The more familiar probability of rejecting the null hypothesis when true ("error of the first kind") is denoted by \( \alpha \) and conventionally is set at .05 or .01. The power of a statistical test is often represented by its "operating characteristic," a curve in which \( \beta \) is plotted as a function of the magnitude of difference between alternative and null hypotheses. Operating characteristics for common statistical tests have been published in refs. 12 and 13.

Suppose that the variance of the individual’s fluctuations \( \Delta \) under the null hypothesis were equal to the average value \( \sigma^2_{A} \). Then, in a two-sided test, the null hypothesis would be rejected however \( (x_t - s_{t-1})/\sigma_A \) \( \geq |z_{\alpha/2}| \), where \( \sigma_H = \sigma_A (w_{t-1} + c + 1)^{1/2} \), and \( z_{\alpha/2} \) is a standard normal deviate corresponding to the probability \( \alpha \). The alternative hypothesis may be represented by the parameter \( \lambda = D/\sigma_H \). The power of the test against an alternative in the direction \( D > 0 \) is equal to the area under the standard gaussian curve from \( (z_{\alpha/2} - \lambda) \) to \( +\infty \). For example, if \( D \) were twice as large as the standard deviation \( \sigma_H \), a two-sided test at the 5% level of significance (\( \alpha = .05 \)) would have a 50/50 chance of rejecting the null hypothesis. To raise this probability to 90%, \( D \) would have to be about 3.3 times as large as \( \sigma_H \). If clinical importance attaches to a change in only one direction, a one-sided test should be used, because this will improve the sensitivity of the test against alternatives in that direction. A one-sided test against \( D > 0 \) at the same level of \( \alpha \) would achieve 90% power when \( \lambda \) reached 2.9.

Suppose that a linear trend of slope \( b \) were present in the variable from the first observation on. Would the normal deviate test detect this nonrandom change in level by, say, the third observation, \( x_3 \)? The answer depends on the "signal-to-noise" ratio of the linear trend, that is, the ratio of the slope \( b \) to the analytic error \( \sigma_A \). The critical value of this ratio (i.e., that value such that any higher value ensures a favorable probability of detecting a linear trend by the third observation) depends on the ratio \( c = (\sigma_A/\sigma_A)^2 \), increasing as \( c \) increases. A mathematical explanation is given in the Appendix.

### A Computer-Generated Example

To obtain data for testing mathematically derived results, I generated 100 sets of four hypothetical serum calcium determinations by random sampling. The sequential observations \( x_1, x_2, x_3, \) and \( x_4 \) in each set were drawn at random from gaussian populations with means \( m_1, m_2, m_3, \), and \( m_4 \) (defined below) and a common standard deviation \( \sigma_A = .068 \) mmol/liter, from the data reported by Bokelund et al. (14) for paired blood duplicates. The ratio \( c = (\sigma_A/\sigma_A)^2 \) was set at 0.6, leading to \( \sigma_A = .62 \) and \( \sigma_A = .53 \). The means were defined as: \( m_1: 2.50; m_2: 2.65 \Delta; m_3: 2.80 \Delta; m_4: 2.95 \Delta \). mmol/liter, respectively, representing a systematic increase of .15 mmol/liter during each time interval, obscured by random shifts \( \Delta \) selected from a gaussian population with mean zero and standard deviation .053 mmol/liter \( [= \sigma_A (c^{1/2})] \). The ratio \( b/\sigma_A \) was .15/0.68, or 2.2, for which the probability of a significant result by the third observation (i.e., \( z_3 = |x_3 - s_3|/\sigma_A (w_2 + c + 1)^{1/2} \), greater than 1.96) was expected to be about 50% (see Appendix). Of the 100 values of \( z_3 \) produced by the simulation, 52 exceeded 1.96, in good agreement with predicted performance.

When the slope was set equal to zero, five of the 100 \( z_3 \)-values exceeded 1.96, in exact agreement with the percentage of false positives expected at the .05 level of significance. Of course, repeated application of the test to successive observations \( x_t, x_{t+1}, \) etc., will increase the probability of at least one false rejection of the null hypothesis with the attendant cost of retesting.

### Using a Graph

Because these statistical testing and estimation procedures are sequential in nature, a graph provides...
a convenient way of accumulating results. The general form of such a graph appears in Figure 2 based on one of the sets of simulated calcium data. Note that the weighted average $s_{t-1}$ is plotted at the abscissal point $t$, along with the observation $x_t$ since $s_{t-1}$ represents the forecast at time $t$. We know that $s_t$ will lie between $s_{t-1}$ and $x_t$. Therefore, if the vertical distance between $s_{t-1}$ and $x_t$ is taken as a unit, then $s_t$ may be found along this distance 100 $w_t$% of the way toward $x_t$, and plotted at time $(t + 1)$. Thus, setting $s_1 = x_1$, $s_t$ may be plotted on the graph without actually calculating its value. Finally, the decision lines $\pm s_{t/2} \sigma_A (w_t + c + 1)^{1/2}$ lie a constant distance above and below $s_t$. The null hypothesis is accepted or rejected depending on whether $x_{t+1}$ falls within or outside these lines (only one line for a one-sided test). Initially, a value of c must be chosen so that $w_t = w_2$ for $t = 2$, and $w_t = w_2$ for $t > 2$ may be obtained from Figure 1 and the distance $z_{t/2} \sigma_A (w_t + c + 1)^{1/2}$ calculated. From then on, only a ruler and graph paper are required for charting the sequence of observations, moving averages, and decision lines.

Discussion and Final Remarks

Population-based norms, even when specified by age and sex, have only limited applicability to any given individual. Nevertheless, they offer valuable guidelines for assessing an individual’s state of health when only single measurements of body constituents are available. When at least two or three serial determinations have been made, the individual has begun to provide his own reference values, and clinical interest now focuses on the significance of observed changes in these values. A statistical theory entirely different from that of “normal ranges” must be called upon to help answer this question. The scheme of analysis becomes necessarily a “bootstrap” operation, because both parameter estimation and testing for significant changes must proceed concurrently. To define precisely what is meant by the statement that the current measurement of Variable X in a patient is significantly different (statistically) from previous measurements, a stochastic model must be constructed incorporating biological and analytical variation. Equations 1 and 2 form, I believe, the most appropriate model which includes these fundamental components of variation.

The methods based on this model may be applied to a combination of several variables, for example, an index of liver or thyroid function derived from a battery of tests. In such a case, the analytical variance would have to be estimated indirectly from a propagation-of-error formula [e.g., Mandel (15)] based on the mathematical form of the index and the variances of the separate tests.

These methods may also be extended to longer series of measurements that are often obtained in clinical research studies, as in intensive drug trials, for example. In such studies, patients may be in control status (i.e., on standard medication) for some time before entering the trial. During this period, the variances of biological fluctuations from time to time can be estimated and analytical variances confirmed. The sequential procedure described here combines simplicity and sensitivity. Moreover, it may easily be incorporated in a computerized system for cumulative storage and reporting of data on patients.

Appendix

1. Power of the normal deviate test to detect a linear trend by the third observation.

Suppose the data consist of at least three equally spaced serial observations containing a linear trend of slope $b$. At $t = 3$, the normal deviate tests the difference between $x_3$ and the weighted average $s_2$, treating both as estimators of $m_3$ under the null hypothesis. Now, $s_2 = w_2x_2 + (1 - w_2)x_1$. Under the alternative hypothesis (trend of slope $b$), $E x_2 = m_3 - b$, and $E x_1 = m_3 - 2b$, so that

$$E s_2 = w_2 (m_3 - b) + (1 - w_2) (m_3 - 2b)$$

$$= m_3 - b (2 - w_2)$$

Since $E x_3$ remains equal to $m_3$, $E (x_3 - s_2) = b (2 - w_2) = D$, the nonrandom change.

Since the denominator of the normal deviate test is $\sigma_H = \sigma_A (w_2 + c + 1)^{1/2}$, the “power” parameter $\lambda$ is given by

$$\lambda = D/\sigma_H = b (2 - w_2)/\sigma_A (w_2 + c + 1)^{1/2}$$

or

$$b/\sigma_A = \lambda (w_2 + c + 1)^{1/2}/(2 - w_2) \quad (5)$$

We noted in the text that if $\lambda = 2$, the normal deviate test would have a 50/50 chance of reaching significance at the 5% level. Therefore, setting $\lambda = 2$ and
evaluating the above equation 5 at a given value of $c$, we can calculate the minimal size of $b/\sigma_A$ required for at least an even chance of obtaining a significant normal deviate test by the third observation. As $c$ increases from 0.25 to 5.0, this critical value of $b/\sigma_A$ increases from 1.87 to 4.55.

References