Sampling from a Skewed Population Distribution As Exemplified by Estimation of the Creatine Kinase Upper Reference Limit

W. Greg Miller,1 Vernon M. Chinchilli,2 Hanns-Dieter Gruemer,1 and Walter E. Nance3

Creatine kinase (EC 2.7.3.2) was measured in sera from 580 females, ages 1–77 years, and 550 males, ages 1–63 years. The distribution of results for male and female groups shows pronounced skewing toward higher values. The observed distribution of results could not be described by any of six mathematical formulas for skewed distributions, an indication of the unsuitability of such formulas to transform these data for parametric analysis. The range of 97.5 percentile estimates produced by six independent samples of 100, 200, and 400 observations randomly selected from a mathematical model defined by the adult female distribution showed progressive narrowing from the 150–380 U/L interval for the samples of 100 observations to 200–265 U/L for the samples of 400 observations; no further improvement was seen when 800 observations were used. The samples of 100 and 200 observations contained extreme value points that might be "outliers" but were shown to be valid members of the population distribution when larger sample sizes were collected.

Additional Keyphrases: Statistics · Reference interval · Nonparametric analyses · Biological vs analytical variation

The activity of creatine kinase (CK; EC 2.7.3.2) in serum is a frequently used marker of cardiac and skeletal muscle disease. The normal reference interval for laboratory results is commonly defined as the central 95% of observations. Thus, for interpretation of results, it is desirable to estimate the 2.5 and 97.5 percentile values with a reasonable degree of certainty. The distribution of CK results for a healthy reference population is markedly skewed toward higher values. Several workers have therefore promulgated various transformations to convert non-gaussian data into a gaussian form from which the 2.5 and 97.5 percentile values can be estimated (1–3). For these transformations to be successful it has been necessary to delete "outlier" values from the data arbitrarily with no medical justification for doing so. Boyd and Lacher (3) were unable to transform one highly skewed model distribution, designated S9(c), which is generally similar to the CK distribution observed here. Furthermore, gaussian transformations used in a study of the multivariate reference range produced severe distortion of individual CK and other test results that were at extremes of the distributions (4).

Reportedly, use of the nonparametric percentile method to estimate the central 95% of a reference population provides results equivalent to those of a parametric method when the data are gaussian, and superior results when the data have been erroneously forced to approximate a gaussian distribution (5, 6). One drawback to nonparametric methods is the requirement for large data bases for reliable estimates of the cutoff values for the central 95% of the population. Here, we examine the data requirements for nonparametric estimation of these cutoff values and the severity of the errors that can result when insufficient data are available to define a skewed distribution adequately.

Materials and Methods

Serum specimens were obtained from an ambulatory normal population by the Department of Human Genetics for a study of twins in twins. That each subject was in good health was assured by a medical history and physical examination. Subjects were not screened for unusual muscular activity before blood was sampled. The specimens were collected over a four-year period from 580 females, ages 1–77 years, and 550 males, ages 1–63 years. When an individual was sampled on more than one occasion, only the first observation was used in this study. Blood specimens were centrifuged within 2 h of collection, and the serum was stored at 4°C and analyzed within two days. No observations were deleted from the data other than those for individuals identified as diseased by the history and physical examination before analysis.

Creatine kinase activity was measured by a modified coupled Rosalki reaction (7) with a kinetic enzyme analyzer (Model KA-150; Perkin Elmer Corp., Norwalk, CT 06856) and reagents from Perkin-Elmer. This instrument provides a 6-min preincubation of serum with buffer, cofactors, and dithiothreitol at 30°C, followed by addition of substrate (overall 20-fold sample dilution), heating to 37.0 ± 0.2°C, and transfer to the measurement cuvet in 120 s. The initial rate of reaction is monitored for 8.8 s, and activity is calculated by use of the molar absorptivity of NADH at 340 nm (6.28 × 104 L/mol cm). The final reaction reagent composition, per liter, is: 0.10 mol of piperezine-N,N'-bis(2-ethanesulfonic acid), pH 6.8, 210 mmol of creatine phosphate, 1.25 mmol of adenosine diphosphate, 170 μmol of glucose, 140 mmol of magnesium chloride, 8 mmol of dithiothreitol, 10 mmol of adenosine monophosphate, 2.25 mmol of NAD+, 8 kU of hexokinase (EC 2.7.1.1), and 4 kU of glucose-6-phosphate dehydrogenase (EC 1.1.1.49). Absorbance and pipettor calibration is performed by quantitative oxidation of NADH to NAD+ coupled with conversion of weighed pyruvate to lactate by the enzyme lactate dehydrogenase (EC 1.1.1.27). Wavelength accuracy is provided by the use of a hollow-cathode lamp and blocking interference filter.

During the early part of this study, i.e., before March 1979, we measured CK activity with the SMAC continuous-flow analyzer (Technicon Instruments Corp., Tarrytown, NY 10591) by a modification of the method of Siegel and Cohen (8). This method involves use of a 21-fold dilution of serum into imidazole buffer (0.14 mol/L, pH 6.8) containing, per liter, 12 mmol of creatine phosphate, 1.5 mmol of adenosine diphosphate, 10 mmol of magnesium chloride, and 5.9 mmol of cysteine. The amount of creatine formed after incubation for approximately 2.4 min at 37°C is measured by dialysis and reaction with alkaline N-ethylmaleimide/diacetyl/orcinol reagent. Endogenous serum interferences are eliminated by use of a blank (the above...
reagents without creatine phosphate). The SMAC method for CK was calibrated by using secondary standards to produce a result equivalent to that produced by the Perkin-Elmer kinetic analyzer (9).

When data collection for this study was terminated, the Perkin-Elmer instrument was replaced by a Cobas-Bio centrifugal analyzer (Roche Analytical Instruments, Nutley, NJ 07119) and the reagent formulation was updated to conform to the recommendations of the Scandinavian Committee on Enzymes (10). Regression analysis by use of the Deming approach (11) between the Cobas-Bio method and the Perkin-Elmer method for 81 patients' specimens (activity range 6–1075 U/L) gave a slope of 1.30 and intercept +4. To apply our results for the normal reference interval to that obtained with the Scandinavian method, we multiplied all results in this study by the regression coefficient 1.30 before statistical analysis. The negligible intercept was ignored. This adjustment permits direct application of reference intervals determined here for results obtained with the optimized and widely used method at 37 °C recommended by the Scandinavian Committee on Enzymes. It has no impact on the statistical analysis of skewed distributions. The exact value of the reference limits in an individual laboratory may differ from the values determined here.

Results

Figure 1 illustrates the stability of the analytical method for CK measurement during the time the healthy population was sampled. The difference between the monthly mean value for lyophilized, pooled-serum quality-control material and the cumulative mean for an entire lot of material is plotted vs the date of analysis. For the KA-150 enzyme analyzer, a maximum difference of 12% and average differences of ±2% were observed from month to month at all three control activities monitored. For the SMAC analyzer, which we used prior to March 1979, average differences of ±5% are seen for the normal control material. The above-normal control material shows a significant decrease in activity during the life of each lot. Considering the consistency of the normal control material and the consistency of the population median (Figure 2), we concluded that the instability was in the quality-control material itself rather than in the analytical method. In a study of 14 quality-control pools used in regional programs during 1973 to 1976, CK activity decreased by an average of 18% per year in 29% of the pools (12).

Figure 2 shows the median CK result for sequential groups of 20 healthy individuals, plotted vs the average date of analysis for each group. The average difference from the overall median varies from 9 to 14% for the three groups. Excepting the single high sub-group median for the adult women, the maximum deviation from the overall median is approximately 30% for each male and female group. There are no noticeable trends in the population studied, which, combined with the analytical stability shown in Figure 1, suggests that laboratory performance was consistent during this four-year period.

Figure 3 depicts the raw data for the distribution of CK activities with age in 580 females and 550 males. The median and 97.5 percentile values for various age groups are also indicated. In females the CK median values decrease during the first two decades of life, but there are no significant changes thereafter until at least age 55. In males there appear to be two distinct groups, below and above age 30, whose medians differ significantly (p < 0.0001) by the Mann–Whitney test.

Figures 4 and 5 show histograms and cumulative frequency plots of CK activities for 379 women, ages 19 and over; 212 men, ages 31 and over; and 338 males, ages 30 and under. The distributions are markedly skewed toward higher values. The data for 379 women could not be described (Kolmogorov–Smirnov p < 0.025) by any of six different mathematical formulas for skewed distributions listed in Appendix A. This confirms the unsuitability of these mathematical distributions to transform the data for parametric analysis in the manner suggested by Boyd and Lacher (4). The effect of the skewed distribution on the interpretation of CK results is illustrated by dividing the range of results that includes 97.5% of the population into quarters. For example, in 379 women, grouped by 60 U/L per quarter, the first through the fourth quarters contain respectively 48%, 39%, 8.5%, and 2% of the population. In this group the 90% confidence interval for the 97.5 percentile value of 240 U/L.

---

**Fig. 1.** Difference between the creatine kinase monthly mean value and the final cumulative mean value for pooled serum quality control material. The arrows indicate changes in lot of material used. For the KA-150 enzyme analyzer, the low-, middle-, and high-activity controls had cumulative mean values ranging from 46 to 71, 141 to 177, and 415 to 636 U/L, respectively. For the SMAC analyzer (used before March 1979) the low- and high-activity controls had cumulative mean values of 64 and 499 to 552, respectively.

**Fig. 2.** Median of creatine kinase results for sequential groups of 20 subjects vs date of analysis.
Fig. 3. Creatine kinase activity vs age for 580 female and 550 male healthy individuals.
Not shown are results for two males, one age 16 with CK = 855 U/L and the other age 30 with CK = 998 U/L. The solid line indicates the median value for the various age subgroups indicated on the top margin. The dashed line shows the uncertainty in estimating the 97.5 percentile value in these small subgroups.

Fig. 4. Number of individuals vs value for creatine kinase activity in three homogeneous subgroups of the population.

Fig. 5. Percent cumulative frequency of results vs value for creatine kinase activity in three homogeneous subgroups of the population.

is 199–333 U/L. The upper limit of this confidence interval excludes four observations, or about 1% of the results. For 338 males 30 years old and younger the 97.5 percentile value is 391 U/L with a 90% confidence interval of 332–549 U/L. The upper limit of this confidence interval excludes three observations, which is 0.9% of the results.

To illustrate the inability of a small sample size adequately to define a skewed population distribution, the data for 379 women was assumed to represent a population with distribution as shown in Figures 4 and 5, and a mathematical model of this distribution was constructed as described in Appendix A. This model distribution was randomly sampled independently six times each to obtain groups of 100, 200, 400, and 800 individual CK results. Cumulative frequency vs CK result for these groups is plotted in Figure 6. The estimate of the 97.5 percentile value varies between 150 and 380 U/L for the groups of 100; 220 and 315 U/L for the groups of 200; 200 and 265 U/L for the groups of 400; and 205 and 300 for the groups of 800 (vs 240 U/L for the model population). The limitations imposed by small samples from a skewed distribution can also be demonstrated by randomly sampling with replacement from the CK results for 379 women. Five hundred samples of 100, 200, 400, and 800 observations each were generated and the 97.5 percentile value for each sample was determined. The mean and standard deviation of the 97.5 percentile values are 259 and
64 U/L for groups of 100; 260 and 57 U/L for groups of 200; 248 and 43 U/L for groups of 400; and 242 and 35 U/L for groups of 800 observations.

Discussion

In this study, no datum on CK activity has been excluded from the pool. Subjects were not screened for unusual muscular activity before blood was sampled. Although serum CK activity can be influenced by exercise (13, 14), it is impossible to quantify such factors as degree of muscle activity or temporal response to stress. Thus this study represents a random sample of healthy individuals.

The criteria for accepting observations into a data set must be established before the data are evaluated and must be based on the health status of the individual. All results satisfying these criteria are equally valid for describing the population distribution. If an individual is subsequently shown to have been diseased at the time the sample was collected, on grounds other than the test result, then it is valid to delete that result. The stability of quality-control results and patients' median values during the time of the study suggests that there is no significant analytical bias in the data. The patients' data were collected over a large time span, which effectively averaged any short-term methodological fluctuations that may have occurred. Because analytical fluctuation is much smaller than the fluctuation of results observed in the healthy population, the range of CK results observed is assumed to reflect biological differences among individuals.

The finding of a highly skewed distribution for creatine kinase is in agreement with reports by other authors (3, 15–18). Because of the small absolute number of observations that fall in the skewed tail region, a large number of observations is necessary to adequately define this area. For example, if a sample of 100 observations were collected for the group of women, only two values would be expected to fall within the fourth quarter of the reference range (180–240 U/L), and two additional values would be expected above the 97.5 percentile value of 240 U/L. Upon casual inspection, these four observations might be classified as "outliers" and be discarded from the data base. Such an arbitrary action would introduce a bias in the assignment of the upper reference limit.

Reiter et al. (5) state that 120 is the minimum sample size that permits the calculation of a 90% confidence interval for the 97.5 percentile value. In this instance, the upper limit for the confidence interval is the largest observation in the sample, designated x(120), where x(m) is the mth ordered observation. Although the confidence interval exists, it may be unreliable if extreme values distort the sample of 120. Therefore, a larger sample would be more appropriate so that the greatest observation is not the upper limit for the 90% confidence interval. As detailed in Appendix B, this upper limit is x(400), when the number of observations (n) equals 400, so that the upper 1% of the ordered sample [x(399) through x(400)] does not influence the confidence interval. As the sample size increases, the effect of extreme values is minimized even further. For example, if n = 800, then the upper limit for the 90% confidence interval is x(788), which eliminates the influence of the upper 1.5% of the ordered sample. Based on this analysis, we recommend that a sample size of approximately 400 be used for adequate protection against extreme values when one is estimating the 97.5 percentile value with a 90% confidence interval.

For CK results for 379 women, the 90% confidence interval (199–333 U/L) includes only 2.4% of the population, although the activity range of 134 U/L is approximately one-half of the entire range of the central 95% reference interval (220 U/L). Furthermore, the confidence interval itself is not symmetrical, the lower limit being 41 U/L below and the upper limit 93 U/L above the 97.5 percentile value of 240 U/L. Evidently, even with sample sizes approaching 400 observations, values in the skewed tail region strongly influence the estimation of the normal reference interval.

Random selection of independent samples of 100, 200, 400, and 800 observations from a model of the distribution defined by our CK results for women demonstrates the unreliability of small numbers of samples in defining a skewed distribution. As expected, the deficiencies are most pronounced in the samples of 100 observations, where each sample defines a markedly different distribution in the skewed tail region. In Figure 6, the three largest observations in a single sample (X in the samples of 100 and □ in the samples of 200) might appear to be "outliers." However, with larger sample sizes it becomes obvious that these points are valid members of the population distribution. As the sample size increases from 100 to 400 observations, the 97.5 percentile value of the six independent samples from the population model covers a progressively narrower range (Figure 6) and the standard deviation of the 97.5 percentile value in 500 independent samples from the population becomes progressively smaller. This reflects the improved ability of larger samples adequately to define the skewed portion of the distribution. Increasing the sample size to 800 observations in these models showed no appreciable improvement over samples of 400 observations. The slightly larger range of 97.5 percentile estimates in Figure 6 may represent the effects of chance on sampling from a population or may be partly ascribed to an anomaly in the random-number generator in the Statistical Analysis System software (SAS Institute, Cary, NC 27511).

The inability to estimate a reliable 97.5 percentile value in small samples from a skewed distribution precludes subdivision of a population sample into smaller groups for this purpose. Figure 3 illustrates the substantial uncertainty in estimating the 97.5 percentile value by age groups in this population. However, the median is robust to the extreme values and can be determined with good reliability even in groups of 20 observations (i.e., Figure 2), providing a good indication of trends in subgroups of a population. Because analytical performance was quite consistent during this study (Figure 1), one can have confidence in any trends shown by the median of subgroups. In contrast to reports by other authors (19, 20), this study shows no trends in CK results for healthy individuals as a function of the time of year at which the sample was collected.

Although a skewed distribution introduces uncertainty in the determination of the 97.5 percentile value, the probability that an individual with a result in the tail region will be in the reference group is less markedly affected by moderate changes in the CK result. Again consider the group of 379 women: a CK value of 240 U/L represents the 97.5 percentile value; however, values of 200 and 160 U/L include 96.5 and 94.5% of the reference population. Consequently, medical decisions based on dichotomization of laboratory results will be significantly affected by reference ranges determined from data that follow a skewed distribution. On the other hand, medical interpretation of the continuum of reference results, such as by using the likelihood ratio technique (27), will maximize the information content available in any reference population distribution. The application of CK results for determination of carrier status in Duchenne muscular dystrophy (22) illustrates sensitivity—specificity relationships and the use of likelihood ratios in highly skewed overlapping populations.
Appendix A

Distribution Functions of Model Skewed Distributions (23, 24)

1. Log normal

\[ F(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{1}{2}t^2} dt \]

2. Log-logistic

\[ F(x) = \left[ 1 + \exp \left\{ -\frac{x-a}{b} \right\} \right]^{-1} \]

3. Extreme value

\[ F(x) = \exp \left\{ -\exp \left\{ -\frac{x-c}{b} \right\} \right\} \]

4. Weibull

\[ F(x) = \exp \left\{ -\left( \frac{x-a}{b} \right)^c \right\} \]

5. Gamma

\[ F(x) = \left\{ \frac{1}{\Gamma(a/b)} \right\} x^{a/b-1} e^{-x/b} \]

6. Generalized gamma

\[ F(x) = \left\{ \frac{1}{\Gamma(a/b)} \right\} x^{a/b-1} e^{-x/b} \]

Because none of these six models adequately described our distribution of CK values on the 379 women, the following mathematical scheme was constructed to model this distribution. The 5 percentile, 10 percentile, \ldots, 90 percentile, and 95 percentile were estimated for CK from our sample. Between any two consecutive percentiles CK was assumed to be uniformly distributed, except that beyond the 95 percentile CK was assumed to be exponentially distributed, to allow for large values of CK. The exponential distribution is a special case of the gamma distribution (when \( \alpha = 1 \)). Random values between zero and one were generated via the Uniform random number generator in the Statistical Analysis System. For example, for a random value of 0.336, the corresponding CK value from the constructed model is found by interpolating between the 30 percentile (CK = 49 U/L) and the 50 percentile (CK = 51 U/L) to yield CK = 50 U/L. When the generated value is larger than 0.95, the appropriate CK value is determined from the exponential distribution.

Appendix B

Estimation and Sample Size Determination for Percentiles

Let \( p \) be a number between zero and one such that \( \hat{p} \) represents the 100p percentile of the population with respect to a particular numerical response. The percentile \( \hat{p} \) is defined as that numerical value in which 100p% of the population is below \( \hat{p} \) and 100(1-p)% is above \( \hat{p} \). For instance, the median is expressed as \( \hat{p} \) and the first quartile \( \hat{p} \).

The population percentile \( \hat{p} \) is rarely known, so it must be estimated from a sample of observations. Let the n observations from the sample be denoted by \( x_1, x_2, \ldots, x_n \), and let the ordered observations be denoted by \( x_{(1)}, x_{(2)}, \ldots, x_{(n)} \). The ordering ranges from smallest to largest in that \( x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)} \). Let the product of \( p \) with \( n+1 \) be \( k+g \) where \( k \) is the integer part \( (1 \leq k \leq n) \) and \( g \) is the fractional part \( (0 \leq g \leq 1) \). Then the sample estimate of the population percentile is

\[ \hat{p} = (1-g) x_k + g x_{k+1} \]

As an example, suppose \( n = 30 \) and \( p = .5 \), which yields \( n+1 = 16.5 \). Then \( \hat{p} = .5 x_{(16)} + .5 x_{(17)} \).

As illustrated by David (25), a distribution-free confidence
interval for $\xi$ can be determined from the ordered observations. This interval estimate of $\xi$ is more informative than the point estimate $\hat{\xi}$. If the desired level of confidence is $(1 - \alpha)100\%$, where $0 < \alpha < 1$, then the lower and upper endpoints of the interval are given by $x_a(r)$ and $x_a(s)$, respectively, where

$$P(x_a(r) \leq \xi \leq x_a(s)) = \sum_{i-r}^{s-1} \binom{n}{i} p(1 - p)^{n-i} \geq 1 - \alpha$$

and $s - r$ is as small as possible.

Regardless of the underlying distribution of the observations, the probability of $x_a(r)$ and $x_a(s)$ covering $\xi$ is the sum of binomial probabilities. Unless a computer routine is available, determining values $r$ and $s$ can be time-consuming. However, the task simplifies somewhat when considering the median. Because of the symmetry with $p = .5$, it is best to take $s = n - r + 1$ so that the above conditions reduce to

$$P(x_a(r) \leq \xi \leq x_{a(n-r+1))} = (.5)^n \sum_{i=r}^{n-r} \binom{n}{i} \geq 1 - \alpha$$

and $r$ as large as possible.

An example for the 95% confidence interval of the median when $n = 20$ follows. For $r = 5, 6,$ and $7$, $P(x_a(r) \leq \xi \leq x_{a(n-r+1)} = .968, .959, \text{and} .885$, respectively. According to the criteria for the median, the appropriate value of $r$ is 6, which yields $[x_{a(6)}, x_{a(15)}]$ as the 95% confidence interval.

A computer routine for calculating the confidence interval of a percentile is written in the Statistical Analysis System (SAS) and is available from the authors. However, if the user does not have access to a computer that supports SAS, other indirect methods for calculating the sum of binomial probabilities are available. For instance, David (25) equates this sum to the difference of incomplete beta functions:

$$\sum_{i=r}^{n-r} \binom{n}{i} p(1 - p)^{n-i} = I_p(r, n-r+1) - I_p(s, n-s+1)$$

where

$$I_p(a, b) = \int_0^p \beta(a, b) - t^{a-1}(1-t)^{b-1}dt.$$

Tables for the incomplete beta can be found in a few mathematical statistics books, such as Pearson and Hartley (26). If the sample size $n$ is very large, then either a Poisson or normal approximation to the binomial can be applied, as discussed by Bhattacharyya and Johnson (27).

This work was supported in part by grants from the Muscular Dystrophy Association to H.D.G. and grants 2P01HD10291 and IR01HD15838 from the National Institute of Health to W.E.N.

References


24. Ibid., 2, p 17.

