Vol. 7 No. 1
Winter 1996

AN AMALGAMATED METHOD FOR ESTIMATING AGE-RELATED CENTILES

S. M. T. Ayatollahi

Department of Biostatistics, Shiraz University of Medical Sciences, Shiraz, Islamic Republic of Iran

Abstract

Monitoring childhood development by use of age-related standards is often hampered in developing countries by the fact that the only available charts are based on inappropriate Western data. Methods for deriving such standards are based on collected data and can easily be applied using locally available expertise and technology, therefore making them of great importance. Two such methods have recently been developed: the LMS method and the HRY method. We compare these methods both in terms of their ability to estimate centiles from several simulated data sets and in terms of their ease of use. The relative merits of the theoretical bases of the methods are also briefly considered. The LMS method essentially worked by applying an age-dependent Box-Cox transformation (the L curve) and then applying Normal theory methods to estimate the mean and coefficient of variation (the M and S curve). The main problem experienced with the method was in its practical application, as the degree of smoothing that should be applied to the L and S curves was sometimes difficult to judge. Different, but apparently reasonable judgements led to markedly different centiles. An application of smoothing techniques used in the HRY method to the problem of smoothing L and S curves was easier to use and produced accurate centiles. A problem with the HRY method is that the degree of non-Normality which this method can accommodate may be limited, as can be shown from its relation to the Cornish-Fisher expansion. However, if the data are first transformed, for example using the L curve from the LMS method, then the range of distribution that the HRY method can accommodate is increased. This is illustrated using simulated data. Although both methods can perform well, the ease and width of their application can be increased by use of techniques used in the other method. It is more realistic and effective if one notes that no medical phenomenon is purely parametric or purely nonparametric. This hypothesis led us to amalgamate the two methods together thus producing more powerful and reliable results than either one could provide individually.

Keywords: Box-Cox power transformation; Growth standards; HRY distribution-free method; LMS parametric method; Smoothed centiles

Introduction

Reference centile curves are used widely in medical practice as a screening tool. They identify subjects who are unusual, in the sense that their value of some particular measurement, for example, height for weight, lies in one or

er tail of the reference distribution. The need for centile ves, rather than a simple reference range, arises when measurement is strongly dependent on some covariate, :n age, so that the reference range changes with the ariate. The case for making centile curves smooth is to ıe extent cosmetic - the centiles being more pleasing to eye when smoothed appropriately - but there is also the lerlying justification that physiologically, small changes he covariate are likely to lead to continuous changes in measurement, so that the centiles ought to change ɔothly. In such cases, fitting discontinuous curves could l to substantial bias.

The literature on fitting smooth centiles to reference a has mushroomed in the last few years [1-9]. Two such thods have recently been developed: the LMS method : to Cole [2] and the HRY method due to Healy, ;bash, and Yang [1]. The LMS method essentially rks by applying an age-dependent Box-Cox [10] ısformation (the L curve) and then applying Normal ɔry methods to estimate the mean and coefficient of iation (the M and S curves). The HRY method is a ıpletely different approach to the problem - namely an irely nonparametric method of centile curve fitting.

No attempt has been made to critically examine the formance of each of the two methods and their problems 'e not been addressed. This led us to compare both thods by their performance on three simulated data sets l resulted in the development of a complementary ɪroach based on amalgamating the techniques used in : or the other method to increase the ease and width of ir application and remove the existing weaknesses of h method. The proposed amalgamated method proved be capable of using both parametric as well as ıparametric aspects of age-related medical data, which ıore realistic than the existing methods.

Methods

The LMS Method

This method, recently developed by Cole, is parametric l relies on Normality. The concept is to use a Box-Cox ver transformation to bring the distribution of the data se to Normality (the L(t) curve), estimate the mean (the t) curve of the transformed data) and standard deviation)) of the transformed data (the S(t) curve) as a function age and hence derive the required centiles. For this son, the method is called the LMS method. Not only :s it provide a coherent set of smoothed centiles with ıtively little compensation, but the shape of the power ve (not to be confused with type II error curve) provides ɔrmation about changing skewness of the distribution ich is not provided by other mehtods of centile fitting -14]. The special feature of this approach is to make the parameters of the transformation a function of age. So, the analysis yields three age-dependent curves.

The variable of interest, denoted by y, is assumed to be positive. Suppose that y has median μ, and that y^{λ} (or if $\lambda = 0$, $\log_e(y)$) is normally distributed. The smoothing can be done using whatever method is convenient, e.g. cubic spline [15], kernel methods [16], polynomials, other specifically mathematical functions [17-18] or simple fitting by eye. It is then appropriate to consider the transformed variable

$$x = [(y/\mu)^{\lambda} - 1]/\lambda, \quad \lambda \neq 0 \qquad (1.1)$$
$$\text{or} \quad x = \log_e(y/\mu), \quad \lambda = 0$$

based on the family of transformations proposed by Box and Cox [10]. This transformation maps the median μ of y to x=0, and is continuous at $\lambda = 0$. For $\lambda = 1$ the standard deviation (SD) of x is exactly the coefficient of variation (CV) of y, and this remains approximately true for all moderate λ. The optimal value of λ is that which minimises the SD of x. The coefficient of variation is relatively independent of the mean.

Denoting the SD of x (and the CV of y) by σ, the SD score (Standard Score) of x (z score) and hence of y is given by

$$z = x/\sigma = [y/\mu)^{\lambda} - 1]/\lambda\sigma, \quad \lambda \neq 0 \qquad (1.2)$$
$$\text{or} \quad z = [\log_e(y/\mu)]/\sigma, \quad \lambda = 0$$

and z has a standard Normal distribution.

Assume now that the distribution of y varies with covariate t, and that λ, μ and σ at t are read off the smooth curves L(t), M(t) and S(t). It follows that

$$z = \{[y/M(t)]^{L(t)} - 1\}/L(t)\,S(t), \quad L(t) \neq 0 \qquad (1.3)$$
$$\text{or} \quad z = \{\log[y/M(t)]\}/S(t), \quad L(t) = 0$$

Rearranging (1.3) shows that the 100α ($0<\alpha<1$) th smoothed centile of y at t is given by

$$C_{100\alpha}(t) = M(t)(1 + L(t)S(t)\,Z_{\alpha})^{1/L(t)}, \quad L(t) \neq 0 \qquad (1.4)$$
$$\text{or} \quad C_{100\alpha}(t) = M(t)\exp[S(t)\,Z_{\alpha}], \quad L(t) = 0$$

where Z_{α} is the normal equivalent deviate (NED) of size α. This shows that if L, M and S curves are smooth, then so are the centile curves.

2. The HRY Method

Another method developed about the same time was aimed at overcoming the problems associated with the traditional methods [11,13]. The method, which is called the HRY method (an acronym composed of the authors' initials), makes no assumption about the nature of the

distribution of the measurements at fixed ages [1].

The method proposed is based on techniques for smoothing a scatter diagram described by Cleveland [21]. Observed measurements are first ordered by age. Then centiles based on counting are found within the narrow age band. Subsequent centiles are found by moving the age band "one point higher", and so on until the whole age range is covered. The raw centiles calculated in this way will be very irregular and need to be smoothed to provide usable centile curves. Not only should each centile follow a smooth curve, but the intervals between centiles at a fixed age should also behave smoothly. The method proposed copes with both these requirements.

To smooth the raw centiles, this nonparametric method first assumes that the 50th measurement centiles can be expressed as a polynomial of degree p (p=1,2,...) in age represented by t. The smoothed value of the 50th measurement centile (Median), M(t), might be:

$$M(t) = b_{00} + b_{10}t + b_{20}t^2 + \ldots + b_{p0}t^p \qquad (2.1)$$

Second, at any given age the other measurement centiles may be expressed in polynomials of the standard normal deviate, z, in relation to the median, i.e.

$$C_p(t) = M(t) + b_0 + b_1 z + b_2 z^2 + b_3 z^3 + \ldots \qquad (2.2)$$

where $C_p(t)$ is the p - th smoothed centile of the measurement (y) and z is the corresponding normal equivalent deviate (NED). In the second equation, we see that if the measurements were exactly normally distributed with standard deviation of $\sigma(y)$, then $b_2=b_3=\ldots=0$ and $b_1=\sigma(y)$. A term in z^2 can account for skewness and in z^3 for kurtosis. The HRY method does not assume that the coefficients $b_0, b_1, b_2, b_3, \ldots$ are fixed but allows them to vary with age (t), so that the whole model, after collecting together terms in t^0, t^1, t^2, etc. from the two equations, may be written as:

$$C_{100\alpha} = (b_{00}+b_{01}z+b_{02}z^2+\ldots)+(b_{10}+b_{11}z+b_{12}z^2+\ldots)t^1+ (b_{20}+b_{21}z+b_{22}z^2+\ldots)t^2+\ldots \qquad (2.3)$$

Since the ages and the z's corresponding to all the raw centiles are known, the model parameters can be calculated by standard least squares (e.g., using any multiple regression procedure). Standard statistical models cannot be used to assess the significance of the model parameters because the raw centile values are all highly correlated as a result of the way they are constructed. The published procedure for fitting this model supposes that measurements are made at a continuum of ages which are not grouped.

The main advantage of this method is the great flexibility in allowing for (a) the centile curves to vary smoothly in a

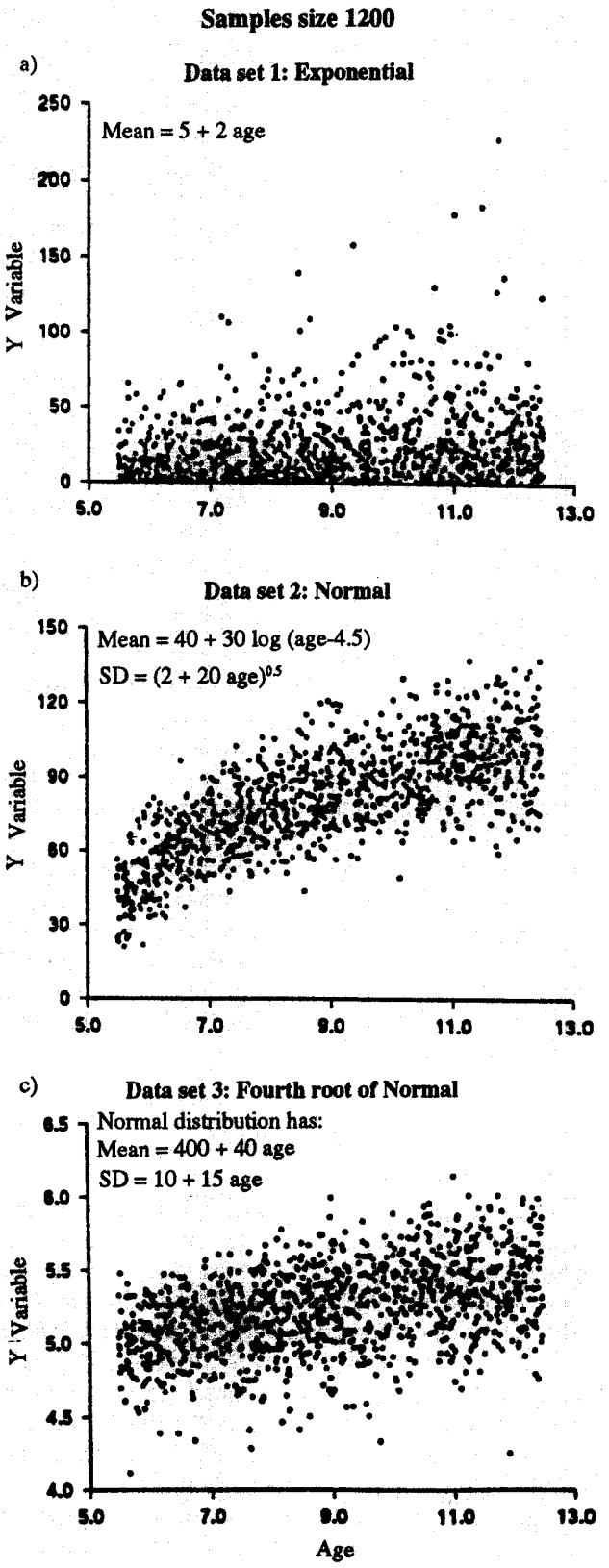

Figure 1. Scatter diagram of Y variable for age for (a) Dataset 1 (b) Dataset 2 and (c) Dataset 3

nonlinear manner with age and (b) non-normal cross-sectional distributions which may change shape, including changing the standard deviation (SD) with age.

Data

Three different data sets were generated by a colleague and centiles fitted by the author, who was totally and absolutely unaware of the distributions used. Once all calculations were made and the centiles fitted, the nature of the data sets was disclosed to the author for judgement and comparison. The first data set was exponential with mean equal to 5+2(age). This was essentially generated to be not Normal and not a power transform of a Normal distribution, which might have been appropriate for the application of the HRY method. The second data set was selected to be Normal but the relation between the mean and age is horrendous. Its mean equals 40+30 [log (age-4.5)] and its variance is equal to 2+20 (age). The third data set followed fourth root of a Normal distribution with mean equal to 400+40 (age) and standard deviation of 10+15(age), as this data set might have been more suitable for LMS method application. Each data set was generated for 1200 points.

Results

We applied the LMS method first and used quarterly, half yearly and yearly age groupings. The results based on half yearly age groupings are presented here in brief. Figures 2a, 2b and 2c present M curves for data sets 1, 2 and 3 respectively. M(t) curves are quite easy to smooth. S(t) curves for the data sets are presented in Figures 3a, 3b, and 3c, and are more scattered but still fairly easy to smooth. Figures 4a, 4b and 4c present L(t) curves for our three data sets. It can be seen that the degree of polynomial fitting and smoothness is difficult to judge. These three curves play the crucial role in smoothing the measurement centiles, among them the role of L(t) curve is the most crucial. The fitted centiles are drawn in Figures 5a, 5b and 5c for the datasets and the true centiles were overlayed on each graph to facilitate comparison. For the sake of clarity, only extreme and median centiles are shown here, but the other common centiles were also calculated.

Comparison of LMS and HRY Methods

Fitting smooth centile curves has always been something of a subjective exercise, or even a black art [2]. The difficulty lies in deciding whether a bump or dip observed on a centile curve at a particular age is a real feature of the data, or whether it is simply sampling error [20].

Both LMS and HRY are powerful and useful methods which produce centiles that are:

1. Smooth
2. Constrained to accord with neighbouring centiles

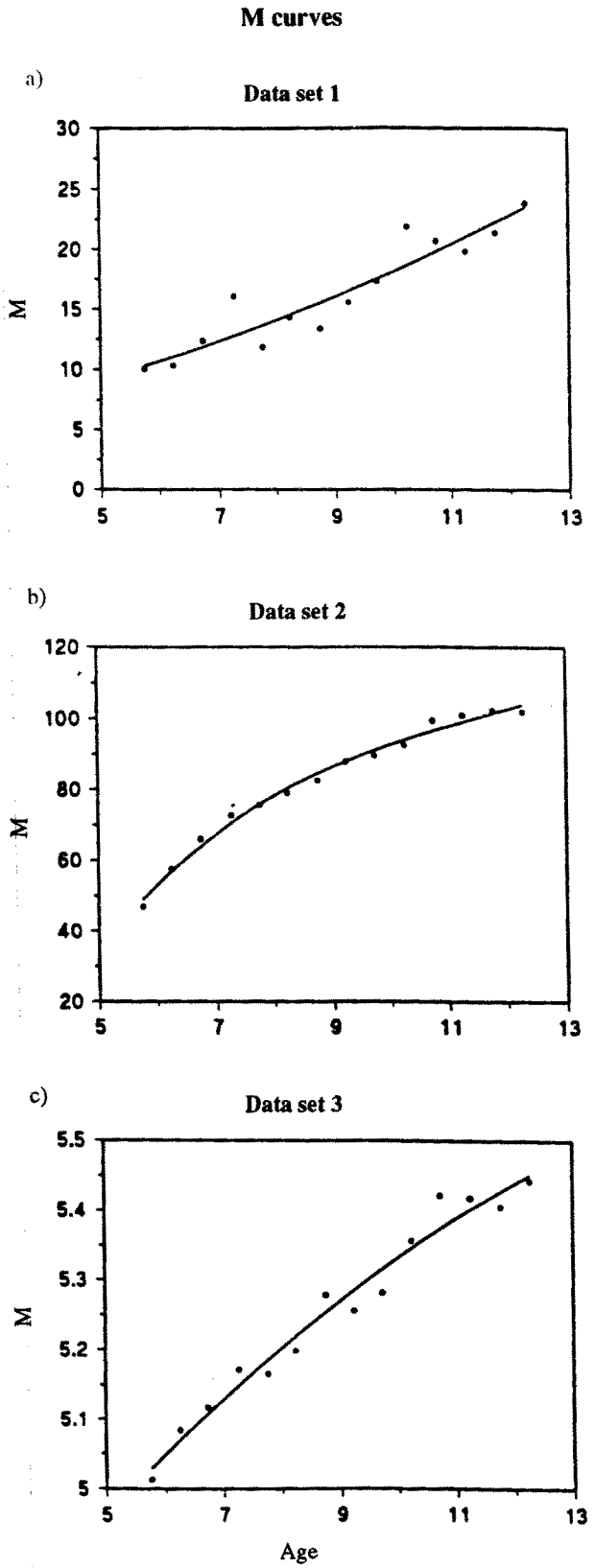

Figure 2. Estimates of the median at each age with the fitted smoothed curve for (a) Dataset 1 (b) Dataset 2 and (c) Dataset 3

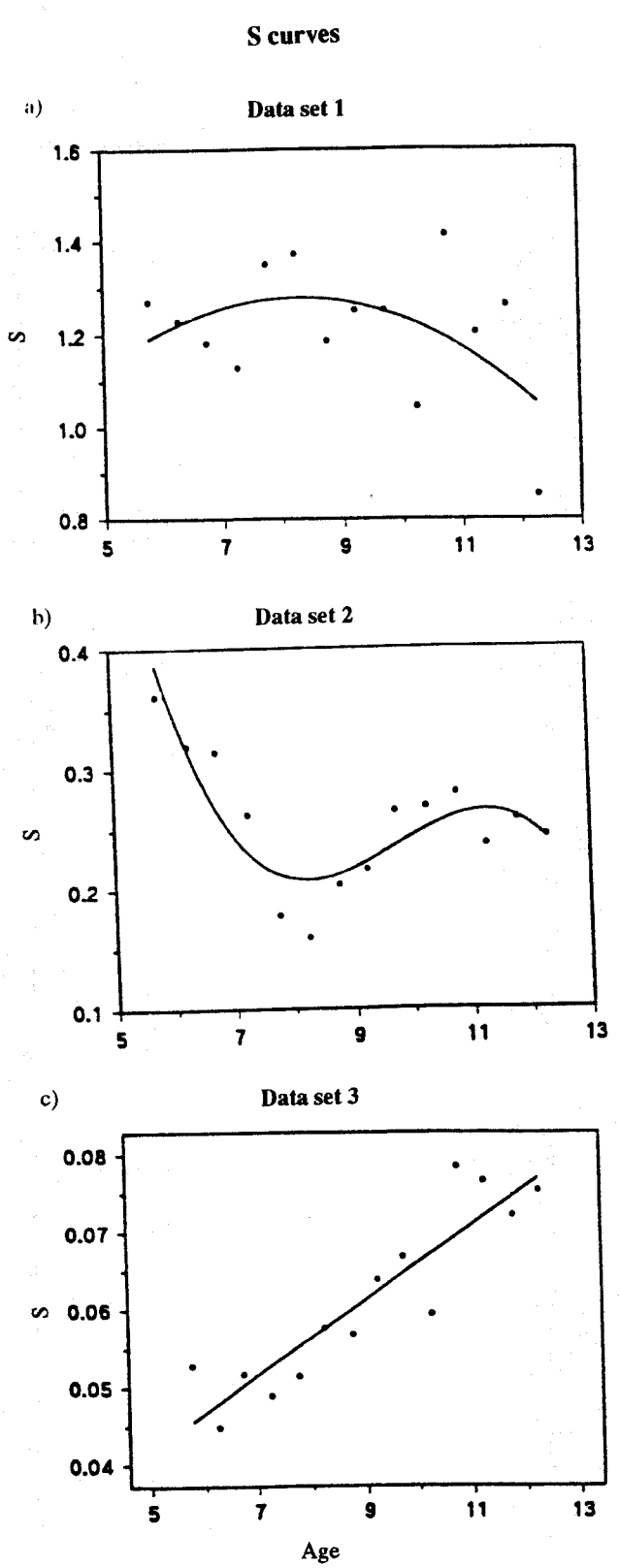

Figure 3. Estimates of the coefficient of variation at each age with the fitted smoothed curve for (a) Dataset 1 (b) Dataset 2 and (c) Dataset 3.

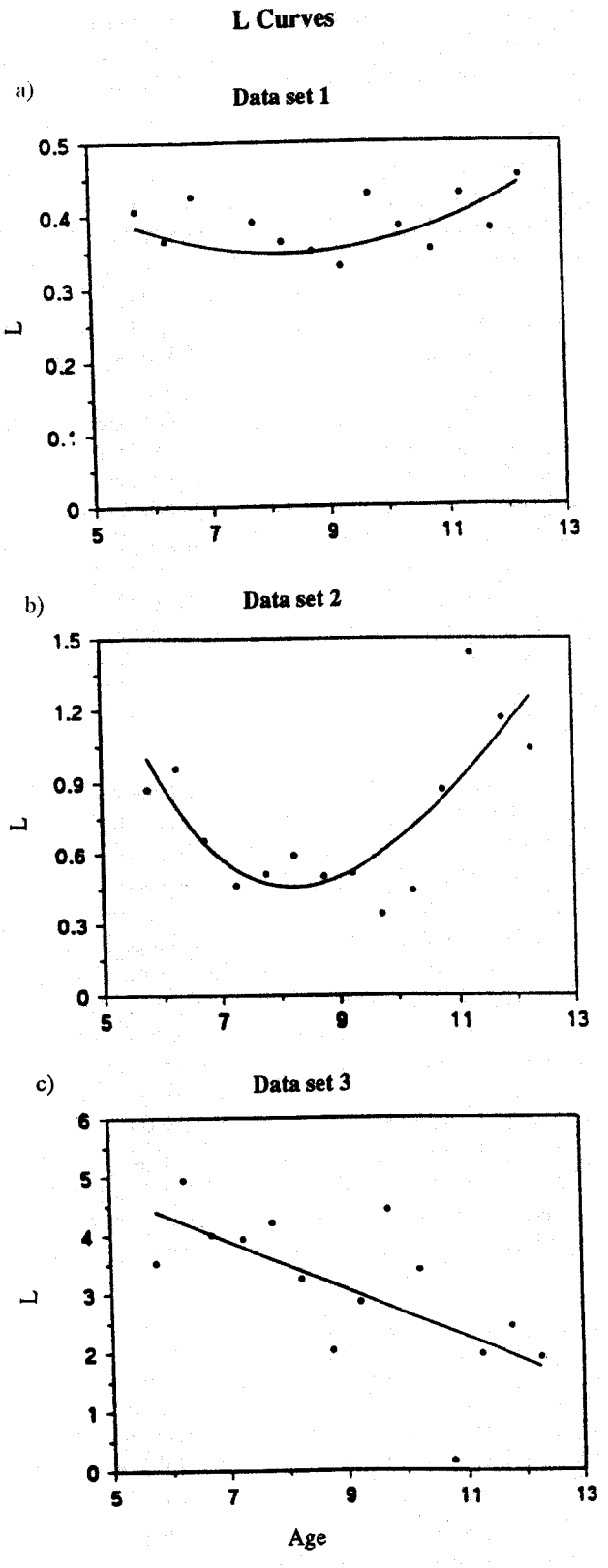

Figure 4. Estimates of the Box-Cox power (the L curve) at each age with the fitted smoothed curve for (a) Dataset 1 (b) Dataset 2 and (c) Dataset 3

and:

3. Allow conversion between z-scores (SD-scores) and centiles.

However, the methods are largely empirical, more so with distributions that are further from Normal.

As age grouping is arbitrary in the LMS method, different results are expected. The narrower the age groupings, the better is the degree of smoothness which results in more sensible centiles. An extension is provided to the method which avoids age grouping and therefore the degree of arbitrariness of the LMS method is decreased to some extent [20]. The amount of smoothing appropriate to L(t) curve gets harder to judge.

Both methods lead to centiles that often appear unnatural and are sensitive to modelling decisions. Therefore, different methods interpret the data differently thus losing their credibility.

Cornish and Fisher [21] showed that for a distribution G(.) that was "close" to Normal, the 100αth centile is approximately

$$+\sigma\{z_\alpha+1/6\rho_3(z^2_\alpha-1)+1/24\rho_4 z_\alpha(z^2_\alpha-3)-1/36\rho^2_3 z_\alpha(2z^2_\alpha-5)$$

that is, the Normal centile $\mu+\sigma z_\alpha$, plus a correction through the standardised third and fourth cumulants, ρ_3,ρ_4 that is polynomial in z. Possible justification of HRY is available through the Cornish-Fisher expansion [22]. Such expansions can only accommodate moderate non-normality, without a large number of terms. The HRY method was later extended to cover wide age ranges [23].

The Amalgamated Method (AM)

These weaknesses have led us to propose a new approach to increase the ease and width of their application. The new approach amalgamates the two methods by applying techniques used in one or the other.

The following algorithm is proposed to simplify use of the amalgamated method:

1. In this step, we first need to calculate λ for each age group or at each age. The resulting λ values are plotted against age. A smooth curve L(t) is then drawn through the points, so that λ's can be read off the curve at any age. The simplest (and computationally most economical) choices of L(t) is when it is a constant, i.e., L(t)=c, which in some cases serves the purpose. Typical values of c are 1,0, and -1 corresponding to the transformations of the measurement itself (measurements are Normally distributed), natural-log (y) and inverse (y), which oversimplifies the procedure. The λ's can be calculated for each age group as described by the LMS method or by using Maximum Likelihood Estimate (MLE). These may cause some difficulties in judging the degree of smoothness of L(t) and as a result make the extreme centiles sensitive and

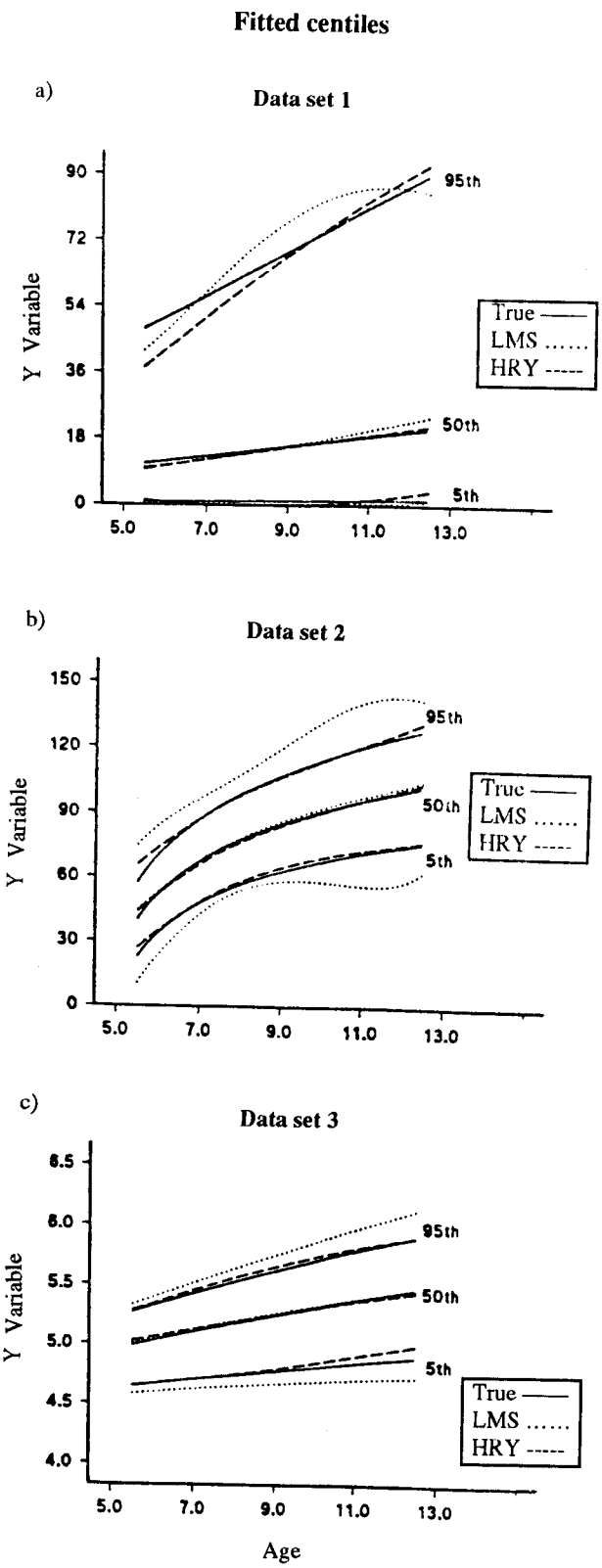

Figure 5. Comparison of the smoothed centiles using the true data, the LMS and the HRY methods for (a) Dataset 1 (b) Dataset 2 and (c) Dataset 3

unreliable.

Therefore, the following technique is proposed to overcome this difficulty. This technique calculates λ's at continuum of ages thus removing the effect of age groupings and hence the degree of arbitrariness of the method. The n observed measurements are first sorted into ascending sequence of age. The first k measurements are then selected, as a working box, where k is an integer and it is a fraction of the total number available, typically 10-20%. The λ for these k measurements is calculated. This procedure has used points 1 to k of the data; it is repeated successively using points 2 to k+1, 3 to k+2,... until the whole span of ages has been covered. Note the age range of these raw centiles will not include approximately k/2 data points at each end of the age-range. These approximately (n-k) λ's are irregular and need to be smoothed. Usually a polynomial fitting is appropriate to use and can model the pattern of λ's effectively giving a smoothed L(t) curve.

2. Apply this transformation to y, i.e. y is transformed to $y^{L(t)}$.

3. Apply HRY to the transformed data, starting from the simplest and lower order polynomials.

4. Undo transformation to get the smoothed centiles.

This new approach was applied to the three data sets and the results are shown in Figures 6a, 6b and 6c. As can be seen, the new amalgamated method produces centiles that are close to the data. Other age groups show similar results which are not presented here for the sake of brevity.

In addition, the box technique for smoothing raw λ's at a continuum of ages was also applied to the simulated data sets. Figures 7a and 7b present application of this procedure and smoothed centiles based on this technique and the amalgamated method for data set 2. Similar results were obtained which are not shown for the sake of brevity.

Discussion

The amalgamated method (AM) takes advantage of the strengths of the LMS and HRY methods and is more powerful than either one singly. The data are nearer to Normal to start with and the exact form of L(t) is much less critical. The new AM method applies two types of Normalisation which are possibly complementary.

Based on a form of mean square error, we were mostly interested in comparing the discrepancies of centiles using each of the three methods with the true centiles. To do so, we arbitrarily considered one of the extreme centiles (95th). The discrepancy was calculated by

$$\text{Discrepancy} = 100\sqrt{[\sum(p(t) - 0.95)^2/n}$$

where p(t) is the true centile of estimated 95th centile at age t. Table 1 gives the discrepancies of using each method with the true values for the three data sets in use. As can be

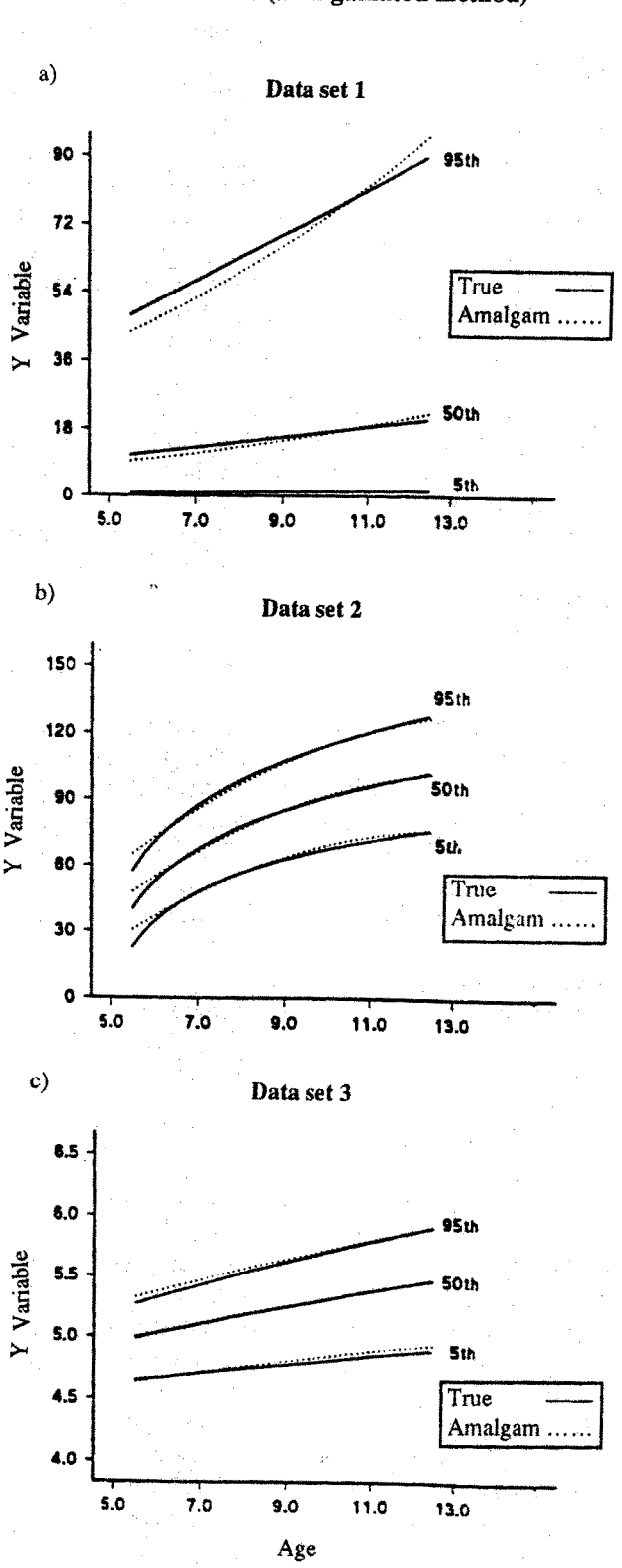

Figure 6. Smoothed fitted centiles using the amalgamated method compared with the true data for (a) Dataset 1 (b) Dataset 2 and (c) Dataset 3

seen from this table, for data set 1 our method produces centiles not much different from those of the LMS and HRY methods. LMS method produces centiles which do not differ greatly from those produced by HRY, but which are completely different from those predicted by the data generator before its application. For data set 2, our method and the HRY method produced the same discrepancies, but the discrepancy associated with the LMS method is awful and is almost four times greater than those of the other two methods. The third data set gave the smallest discrepancy with the HRY method, but it was not substantially different from what has been found using our method. However, the LMS method produced centiles which were very different from the true values, giving results which differed by almost threefold from the other two methods. This was also contrary to our expectations, as data set 3 is a power transform of a Normal distribution and was expected to transform to Normal easily. Similar results were obtained for the lower extreme centile (5th centile) which showed an advantage of the amalgamated method over the other two methods under discussion.

In summary, both the LMS and HRY methods perform well on data whose distribution is close to Normal. Practical problems emerge with data that are not close to Normal. More specifically, in the LMS method the L curve is more difficult to smooth, and with the HRY method uncomfortably high order polynomials are needed for coefficients, which is a notoriously tricky problem and some care is required in using it.

The amalgamated method (AM) proposed here eases problems and improves estimation. Applications of this method to some real cross-sectional and longitudinal growth data are presented elsewhere.

Computing

The computing demands of the AM method are within the range of a moderately powerful microcomputer. It has been implemented using FORTRAN programmes, SPSS [24], MINITAB [25], NANOSTAT, and GROSTAT [26] softwares and called AMSTAT. This provides the user with the specialised commands for age-related centile curves and can handle both distance and velocity data. AMSTAT is specifically written in a way that makes it versatile and capable of being used safely by those who have received only minimum training and thus lack expertise. Further information on AMSTAT can be obtained from the author. Also, the graphs of this paper have all been drawn using Fig. P. graphics software [27].

Acknowledgements

This work was partially supported by grant number 73-170 of the research committee, Shiraz University of Medical Sciences.

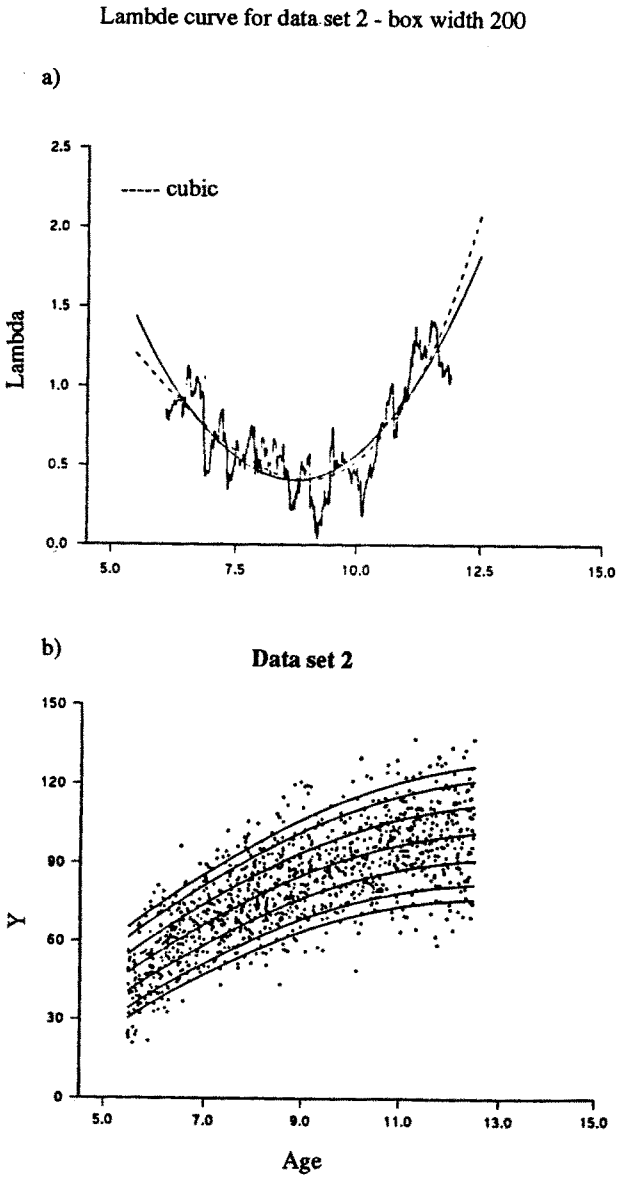

Figure 7. (a) Raw and smoothed fitted curve of the L(t) applying the proposed box moving technique for dataset 2
(b) Scatter diagram and the smoothed centiles for dataset 2 using the amalgamated method

Table 1. Discrepancy of the 95th centiles produced by LMS, HRY and the amalgamated methods with the true centiles using the three simulated data sets

Data	LMS	HRY	Amalgamated
Data set 1	1.1	1.6	0.9
Data set 2	4.6	1.2	1.2
Data set 3	3.9	1.2	1.4

References

1. Healy, M.J.R., Rasbash, J. and Yang, M. Distribution-free estimation of age-related centiles. *Annals of Human Biology*, **15**, 17-22, (1988).
2. Cole, T.J. Fitting smoothed centile curves to reference data (with discussion). *Journal of Royal Statistical Society*, Series A, **151**, 385-418, (1988).
3. Green, P.J. In the Discussion of Cole (1988). *Ibid.*, 410-411, (1988).
4. Jones, M.C. *Ibid.*, 412-413, (1988).
5. Jones, M.C. and Hall, P. Mean squared error properties of kernel estimates of regression quantiles. *Statistics and Probability Letters*, **10**, 283-289, (1990).
6. Pan, H.O., Goldstein, H. and Yang, Q. Nonparametric estimation of age- related centiles over wide age ranges. *Annals of Human Biology*, **17**, 475-481, (1990).
7. Thompson, M.L. and Theron, G.B. Maximum likelihood estimation of reference centiles. *Statistics in Medicine*, **9**, 539-548, (1990).
8. Cole, T.J. The LMS method for constructing normalised growth standards. *European Journal of Clinical Nutrition*, **44**, 45-60, (1990).
9. Rossiter, J.E. Calculating centile curves using kernel density estimation methods with application to infant kidney lengths. *Statistics in Medicine*, **11**, 1693-1701, (1991).
10. Box, G.E.P. and Cox, D.R. An analysis of transformations. *Journal of the Royal Statistical Society*, Series B, **26**, 211-252, (1964).
11. Healy, M.J.R. Notes on the statistics of growth standards. *Annals of Human Biology*, **1**, 41-46 (1974).
12. Healy, J. R. Statistics of growth standards. In *Human growth*, (ed. F. Falkner and J.M. Tanner), pp. 47-58. London: Ballierie Tindall, 2nd edition, (1986).
13. Healy, M.J.R. Growth curves and growth standards: the state of art, Auxology 88. Perspectives in the science of growth and development (ed. J.M. Tanner), London: Smith-Gordon, (1989).
14. Van't Hof, M.A., Wit, J.M. and Roede, M.J. A method to construct age references for skewed skinfolds data using Box-Cox transformations to normality. *Human Biology*, **57**, 131-139, (1985).
15. Silverman, B.W. Some aspects of the spline smoothing approach to non-parametric regression curve fitting. *Journal of Royal Statistical Society*, B, **47**, 1-52, (1985).
16. Gasser, T., Muller, H.G., Kohler, W., Molinari, L. and Prader, A. Nonparametric regression analysis of growth curves. *Annals of Statistics*, **12**, 210-229, (1984).
17. Jenss, R.M. and Bayley, N. A mathematical method for studying growth in children. *Human Biology*, **9**, 556-563, (1937).
18. Preece, M.A. and Baines, M.J. A new family of mathematical models describing the human growth curve. *Annals of Human Biology*, **5**, 1-24, (1978).
19. Cleveland, W.S. Robust locally weighted regression and smoothing scatter plots. *Journal of the American Statistical Association*, **79**, 829-836, (1979).
20. Cole, T.J. and Green, P.J. Smoothing reference centile curves: The LMS method and penalized likelihood. *Statistics in Medicine*, **11**, 1305-1319, (1992).
21. Cornish, E.A. and Fisher, R.A. Moments and cumulants in the specification of distributions. *Revue Institute Internationale de Statistique*, **5**, 307-320, (1937).
22. Matthews, J.N.S. and Cole, M. Cornish-Fisher expansion and calculation of centiles for nearly Normal data. Manuscript: Department of Medical Statistics, Newcastle University, (1994).
23. Pan, H.Q., Goldstein, H. and Yang, Q. Non-parametric estimation of age-related centiles over wide age ranges. *Annals of Human Biology*, **17**, 475-481, (1990).
24. SPSS\PC: Statistical Package for the Social Sciences. Version 6.0. Chicago: SPSS Incorporation, (1994).
25. MINITAB For Windows. Minitab Release 9.2. PA: Minitab Incorporation, (1993).
26. GROSTAT. A Programme for Estimating Age-Related Distribution Centiles. London School of Hygiene and Tropical Medicine, (1988).
27. FIG. P: The Scientific Fig. Processor, Version 6.0, Cambridge: BIOSOFT, (1991).