STATISTICS IN MEDICINE
Statist. Med. 2000; **19**:697–713

Estimating age-related trends in cross-sectional studies using S-distributions

A. Sorribas[1,*,†], J. March[1] and E. O. Voit[2]

[1] *Departament de Ciències Mèdiques Bàsiques, Universitat de Lleida, Av. Rovira Roure 44, 25198-Lleida, Spain*
[2] *Department of Biometry and Epidemiology, Medical University of South Carolina, Charleston, SC, 29425, U.S.A.*

SUMMARY

Growth trends in children are often based on cross-sectional studies, in which a sample of the population is investigated at one given point in time. Estimating age-related percentiles in such studies involves fitting data distributions, each of which is specific for one age group, and a subsequent smoothing of the percentile curves. The first requirement for this process is the selection of a distributional form that is expected to be consistent with the observed data. If a goodness-of-fit test reveals significant discrepancies between the data and the best-fitting member of this distributional form, an alternative distribution must be found. In practice, there is seldom an objective argument for selecting any particular distribution. Also, different distributions can yield very similar fits, so that any selection is somewhat arbitrary. Finally, the shapes of the observed distributions may change throughout the age range so drastically that no single traditional distribution can fit them all in a satisfactory manner. To overcome these difficulties in population studies, non-parametric smoothing techniques and normalizing transformations have been used to derive percentile curves. In this paper we present an alternative strategy in the form of a flexible parametric family of statistical distributions: the *S-distribution*. We suggest a method that guides the search for well-fitting S-distributions for groups of observed distributions. The method is first tested with simulated data sets and subsequently applied to actual weight distributions of girls of different ages. As far as the results can be tested, they are consistent with observations and with results from other methods. Copyright © 2000 John Wiley & Sons, Ltd.

1. INTRODUCTION

The assessment of growth trends in populations of children is often based on cross-sectional studies, in which the population is investigated at one given point in time [1]. The main goal of such studies is the determination of age-specific reference intervals that can be used for screening purposes and for a general characterization of the children's health status. The reference intervals

* Correspondence to: A. Sorribas, Departament de Ciències Mèdiques Bàsiques, Universitat de Lleida, Av. Rovira Roure 44, 25198-Lleida, Spain
† E-mail: Albert.Sorribas@cmb.udl.es

Contract/grant sponsor: CICYT; contract/grant number: PM099-99
Contract/grant sponsor: La Paeria

CCC 0277–6715/2000/050697–17$17.50

Received October 1998
Accepted May 1999

are usually defined as symmetric percentiles around the median. When the measurement of interest varies with age, it is common practice to estimate smooth curves showing the trend of the reference intervals with age. A smooth trend reflects that the change in distribution is progressive with age and that consecutive age groups exhibit similar distributions. Given sufficient smoothness, interpolations can be made for any desired age.

Age-related percentiles are currently computed by a number of procedures that are based on one of two strategies [2]: (i) interpolation techniques without assumptions about the underlying statistical distributions; and (ii) parametric methods based on the normal distribution or on a suitable, normalizing transformation of the data. Among the interpolation techniques, the method suggested by Healy *et al.* [3] and extended by Pan *et al.* [4] has received much attention. This method involves several steps. First one ranks the data by age and forms a group from the first n individuals, where n is at least 50. Individuals 2 to $n + 1$ form a second group, and further groups are composed from individuals 3 to $n + 2$, 4 to $n + 4$, and so forth, until the entire data range is exhausted. In the second step, 'raw' percentiles are estimated for each group, and the totality of these estimates is fitted with a polynomial of sufficiently high order. Finally, the coefficients of the polynomials are smoothed by regressing them on the normal equivalent deviates of the percentiles. Since no underlying distributions need to be estimated, this technique provides an attractive characterization of the percentile curves. Indeed, many 'standard growth curves' used in public health are based on this type of procedure [5]. An alternative to Healy's technique is the method suggested by Tango [6], which uses *smoothed additivity and variance stabilization*. The strength of these methods is also their main disadvantage; since no parametric estimation is involved, the result is in a form that sometimes makes further numerical characterizations and comparisons difficult.

Parametric methods for estimating age-related reference ranges are typically based on the normal distribution [7] and on polynomial fitting of the sample data [8]. Since the age-conditional distributions are seldom normal, Wright and Royston [9] suggested a relatively simple method based on an exponential transformation of the standardized variable. This transformation leads to a distribution with three parameters, which correspond to the median, scale and skewness. Fractional polynomials in age are fitted to these parameters and yield the desired description of the data, and subsequently, estimates for percentiles.

The *LMS method* is an alternative parametric method. It was developed by Cole and Green [10] and assumes that a Box–Cox power transformation [11] normalizes the original variable. Upon normalization, penalized likelihood estimation optimizes three natural cubic splines that model the age-dependent trends in: (i) the Box–Cox power λ used in the transformation (*curve L*); (ii) the mean (*curve M*); and (iii) the coefficient of variation (*curve S*). Application of this method to the estimation of reference percentiles for different measurements in a human population can be found in Cole *et al.* [1].

Some data of interest form age-dependent distributions that no single traditional distribution can adequately model without prior transformation of the data. A case in point is a study relating birth weights to weeks of gestation, in which the authors report skewness values ranging from 0.05 to 7.30 [2]. Studying the weights of girls at different ages, Cole *et al.* [1] identified a trend from normality at birth, to a log transform at 1 year, to an inverse transform at 9 years, and back to a log transform at 14 years. The authors were able to model these different shapes with a continuous trend in the Box–Cox power, but it is interesting, nevertheless, that a representation without prior data transformation would require repeated switches from one distributional form to another. In some time trends, the changes in distributional shape are even more extreme. For

instance, size distributions of even-aged trees are known not only to change the weights of their tails, but even to reverse skewness [12–14].

It would clearly be an advantage if one could find a single distributional form covering the entire spectrum of observed distributional shapes. As Cole *et al.* [1] pointed out, a functional form represents summary curves in a parsimonious fashion and allows the assessment of asymptotic behaviour. Furthermore, an explicit functional form facilitates comparisons, hypothesis testing, and classification. A potential limitation of a single functional form is insufficient flexibility in modelling the data distributions in all age classes. This lack of flexibility could theoretically be addressed by the construction of some 'superfamily' of distributions that would contain all traditional distributions of interest as special cases. To some degree, efforts in this direction, beginning over a hundred years ago with Pearson [15] (see also Johnson and Kotz [16]), have had some success. However, all-encompassing superfamilies, such as the one proposed by Savageau [17] and extended by Voit and Rust [18], are so unwieldy that they are of very limited use in practical data analyses.

In this paper, we propose a parametric method for estimating trends in distributions that is based on the *S-distribution* [19]. With a few trivial exceptions, this distribution does not contain the traditional distributions as exact special cases, but it does approximate very many distributions with high accuracy. This feature of approximately subsuming distributions with different shapes and types of skewness had led to novel classifications of continuous and discrete distributions [19–21] and to statistical analyses, for instance in environmental health risk assessment, that would otherwise have been cumbersome [22, 23].

Our goal is to demonstrate the ability of the S-distribution in providing accurate representations of trends in age-dependent distributions. First, we discuss the quality of data fits with the S-distribution in comparison with some traditional distributions. Secondly, we characterize trends, showing that the S-distribution provides good fits to entire sets of data and to accurate descriptions of the observed trends. Finally, we assess the performance of the proposed technique using simulated data sets and actual data from a cross-sectional study on the growth of Spanish children.

The conceptual components of the proposed method are quite similar to those of the *LMS* method [10]: a parametric distribution forms the basis for data modelling, and the age trends are characterized by polynomials in parameters, which allow for more or less smoothness in the percentile curves. However, there are also crucial differences. The proposed method does not transform the original data, there is not assumption of normality for the transformed data, and the ultimate result consists of a smooth family of parametric distributions in the original age-dependent random variable. Not just the distributions of subsequent age classes have a natural 'commonality', as it is considered desirable in the pertinent literature (for example, see discussion of methods in the paper of Cole and Green [10]), but *all* distributions constituting the resulting distribution family are structurally equivalent and simply differ smoothly in their parameter values.

2. REPRESENTATION OF OBSERVED DISTRIBUTIONS

2.1. Fitting traditional distributions to data

The selection of a distribution for fitting data is not a trivial problem. Very rarely are there theoretical reasons for selecting a particular distribution, and the selection is thus subject to

arbitrariness and possibly a sequence of trials and errors. Even worse, simulation studies show that objective criteria, such as the minimal residual error, do not necessarily yield the true underlying distribution.

As a demonstration of this surprising fact, we fit different distributions to simulated data sets generated from normal, Weibull and gamma distributions, using as the optimization criterion the usual sum of squared errors, SSE. For each of the three different distributions, we fit 25 samples with 160 data points each. As Table I clearly shows, the true distribution, from which the sample was obtained, does not necessarily produce the best fit, even though the sample size is fairly large. For example, data generated from a normal are actually best modelled by a normal distribution only in a third of the cases, while for two-thirds of the cases, the log-normal, Weibull or gamma distribution produces the best fit. The reason for these 'misclassifications' is that, within the stochasticity of the sample, different distributions can have essentially the same shape and yield equivalent fits if their parameters are specified accordingly.

In these simulated scenarios, the true underlying distributions are known. Of course, this is not the case in an analysis of real data, and one might have to optimize several distribution families before finding the best possible fit. As an actual example, consider the weights of girls at different ages, fitted with some standard distributions (Table II). Not only do the different distributions often produce fits of comparable quality, the best-fitting distributional type changes from one age group to the next. The Gumbel distribution seems to be the most appropriate distribution for many age classes, although in some classes the log-normal, logistic, or gamma yields a lower SSE.

2.2. S-distribution: basic concepts and approximation of traditional statistical distributions

The S-distribution was introduced a few years ago as a distribution that can model a wide range of shapes and all types of skewness. It is given as the solution of a differential equation in so-called *S-system form* [19, 20]. The variable of interest in this equation is the cumulative distribution function (CDF), F, and the differentiation is executed with respect to the random variable X:

$$\frac{\mathrm{d}F}{\mathrm{d}X} = \alpha(F^g - F^h) \tag{1}$$

Counting the initial condition $F(X_0) = F_0$, which determines its location, the S-distribution has four parameters. As the value of a CDF, F_0 lies between 0 and 1, and the remaining parameters satisfy the conditions $\alpha > 0$ and $g < h$. The non-linear differential equation in equation (1) has explicit closed-form solutions only for some special cases (however, see Voit and Savageau [24]). For instance, one obtains the exponential distribution for $g = 0$ and $h = 1$ and the logistic distribution for $g = 1$ and $h = 2$. For most other cases, the solution is obtained numerically.

Since the probability density function (PDF), f, equals the derivative of F, the S-distribution can be written in purely algebraic form as

$$f = \alpha(F^g - F^h) \tag{2}$$

The hallmark of the S-distribution is its relative simplicity, combined with its flexibility in shape (Figure 1). With an appropriate choice of parameter values, it rather accurately models most traditional continuous – and even discrete – distributions [19, 20]. Given a traditional distribution, the corresponding S-distribution is obtained by fitting a set of points $(F(X_i), f(X_i))$ with some non-linear fitting routine. For instance, a Gumbel distribution with parameters $\eta = 35$, $\tau = 6$ is represented by an S-distribution with parameters $\alpha = 1.031$, g $= 0.922$, $h = 1.084$, and the

Table I. Artificial data sets are fitted with different distributions, namely, a normal with mean μ and standard deviation σ, a Weibull with PDF $f(X) = \alpha\beta^{-\alpha}X^{\alpha-1}\exp(-(X/\beta)^{\alpha})$, and a gamma distribution with PDF $f(X) = \beta^{-\alpha}X^{\alpha-1}\exp(-X/\beta)/\Gamma(\alpha)$. Parameter values were selected so that the resulting Weibull and gamma distributions are close to symmetric. In each case, 25 samples were generated with 160 data points each. Each sample was subsequently fitted with different distributions, and it was recorded how often each distribution yielded the lowest sum of residual errors, SSE.

	Normal	Gumbel	Log-normal	Logistic	Weibull	Gamma
Normal $\mu = 100$, $\sigma = 6$	32%	0%	28%	16%	16%	8%
Weibull $\alpha = 3$, $\beta = 100$	36%	0%	4%	24%	24%	12%
Gamma $\alpha = 100$, $\beta = 1$	20%	4%	24%	20%	12%	20%

Table II. Sums of residual errors, SSE for different distributions, fitted to 2631 weight data of girls at different ages. Bold type indicates the best-fitting distribution for each age group. Data from Puente *et al.* [5].

Age	Mean	SD	*n*	Normal	Gumbel	Log-normal	Logistic	Weibull	Gamma
5.5	20.5	3.6	109	0.038	0.029	**0.021**	0.037	0.084	0.025
6.	21.8	3.6	87	0.091	**0.050**	0.059	0.081	0.167	0.068
6.5	23.5	4.2	91	0.116	**0.039**	0.073	0.102	0.197	0.086
7.	23.2	4.1	87	0.202	**0.065**	0.130	0.186	0.305	0.153
7.5	26.4	4.9	114	0.131	**0.033**	0.066	0.133	0.222	0.084
8.	27.2	4.7	127	**0.010**	0.025	0.054	0.095	0.207	0.067
8.5	28.9	5.4	108	0.214	**0.067**	0.129	0.205	0.322	0.155
9.	30.3	5.2	111	0.066	**0.027**	0.028	0.064	0.139	0.037
9.5	31.9	5.5	80	0.069	**0.028**	0.037	0.071	0.119	0.045
10.	33.4	7.0	75	0.051	**0.029**	**0.029**	0.052	0.091	0.034
10.5	35.3	7.0	120	0.106	**0.047**	0.055	0.119	0.161	0.067
11.	38.1	7.7	161	0.050	0.045	**0.027**	0.062	0.105	0.030
11.5	39.3	8.0	102	0.080	0.094	0.071	**0.068**	0.120	0.070
12.	43.6	8.6	67	0.037	0.040	**0.029**	0.045	0.054	0.029
12.5	45.6	9.3	120	0.033	0.025	**0.009**	0.033	0.080	0.011
13.	48.8	8.4	129	0.031	0.072	0.035	0.037	0.058	**0.030**
13.5	51.1	8.6	126	0.030	0.065	0.028	0.028	0.072	**0.026**
14.	52.3	7.8	122	0.076	0.076	0.060	**0.058**	0.146	0.063
14.5	52.5	7.9	128	0.130	**0.051**	0.090	0.110	0.242	0.102
15.	54.6	9.1	107	0.064	**0.032**	0.033	0.068	0.129	0.040
15.5	54.9	9.7	110	0.083	0.078	0.068	**0.064**	0.145	0.071
16.	54.1	9.7	114	0.075	**0.033**	0.042	0.081	0.139	0.050
16.5	55.0	6.8	133	0.097	**0.046**	0.063	0.088	0.223	0.073
17.	56.6	8.0	92	0.098	**0.025**	0.060	0.094	0.190	0.072

PDF and CDF of the two are essentially indistinguishable. Other examples are found in the literature [19, 20].

It is worth nothing that the S-distribution preserves relationships between different traditional distributions, such as the approach of the normal by *t* and gamma distributions, if the characteristic parameters tend toward infinity [19].

The fact that traditional distributions can be validly approximated by S-distributions implies that data fits with the S-distribution have SSEs comparable to the best-fitting traditional

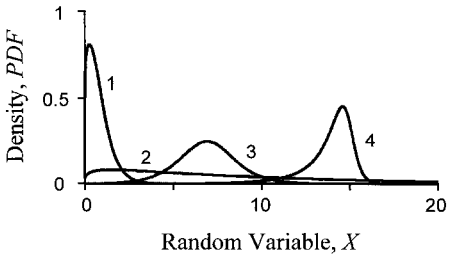

Figure 1. Flexibility of the S-distribution. Densities (PDFs) of S-distributions with parameters: 1 $g = 0.1$, $h = 2$; 2, $g = 0.4$, $h = 0.5$; 3, $g = 1$, $h = 2$; 4, $g = 1.2$, $h = 4.5$. In all cases $\alpha = 1$ and $F(0) = 0.001$.

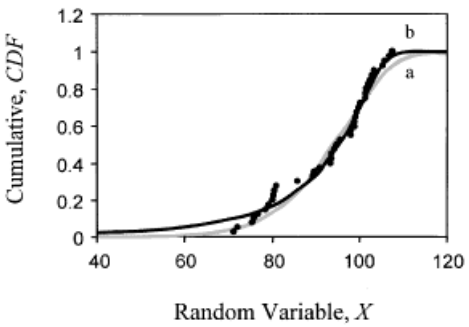

Figure 2. Fitting a sample of Weibull distributed data with an S-distribution and with a Weibull distribution. Data are generated from a Weibull distribution with $\alpha = 10$ and $\beta = 100$. Sample size is 40. Grey line: fitted Weibull ($\alpha = 9.59$, $\beta = 97.71$), with residual error SSE = 0.118. Black line: S-distribution ($g = 0.081$, $h = 1.13$, $\alpha = 7.86$, F (95.35) = 0.5) with residual error SSE = 0.041.

distributions. Furthermore, since the S-distribution also allows for combinations of parameter values that do not correspond to traditional distributions, one might in many cases expect an even better fit than is possible with the traditional distributions. Figure 2 exhibits such a case. Data were generated from a Weibull distribution with parameters $\alpha = 10$, $\beta = 100$, and subsequently fitted with a Weibull distribution and with an S-distribution. While the Weibull distribution is fairly well estimated and yields an acceptable SSE and a satisfactory graphical fit to most of the data, the S-distribution actually returns a visually better fit as well as a lower SSE. The lower SSE *per se* does not imply that the S-distribution is necessarily a better model for these particular data, but that it can serve as a valid default if the actual distribution is unknown. If data are skewed to the left, distributions like the gamma and the log-normal are unsuitable, yet the S-distribution is still able to capture such a shape.

2.3. S-distribution fitting of observed data

The fitting of an S-distribution to observed data can be accomplished through integration of (1) and minimization of the corresponding SSE using the sample CDF as data. We developed a specially tailored module in Mathematica© for this purpose and used it for all analyses of

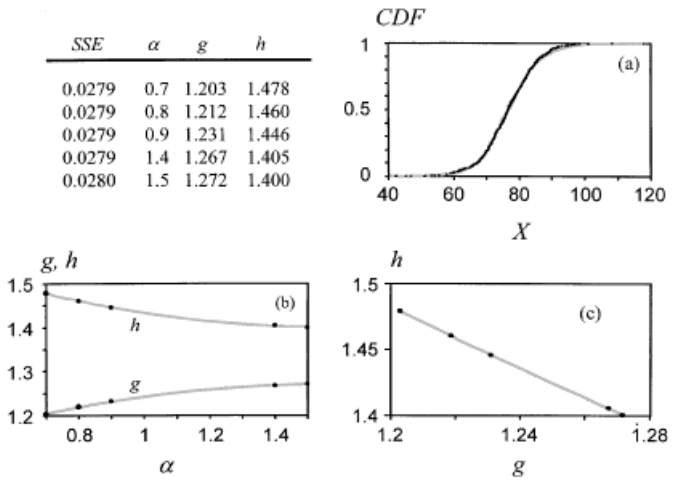

Figure 3. Quasi-equivalent S-distributions. For a given set of data, one may obtain slightly different optimized S-distributions for different pre-set α values. The resulting S-distribution parameters and residual errors are shown in the table. The parameter values of the five sets form a well-defined pattern in parameter space. (*a*) Dots show the sample cumulative. The five fitted S-distributions curves are indistinguishable within the accuracy of the graph; they are represented by the grey line. (*b*) Trends in the fitted parameters g and h as functions of α. (*c*) Trend in the fitted parameter h as a function of g, given α from panel (*b*).

this study. To fit an S-distribution to a set of data, the module computes in the first step the median and uses it to initiate the differential equation solver with $F(\text{median}) = 0.5$. In the second step, the minimization routine `FindMinimum`, which is included in the standard Mathematica© language, determines the optimal triplet (α, g, h) from a user-supplied initial guess.

A common problem is that the minimization surface is relatively flat over a wide region of the parameter space, which makes it difficult to locate a global minimum. Expressed differently, different sets of parameter values lead to data fits of almost identical quality. This type of redundancy is the price for the high flexibility of S-distributions. Besides the logistic difficulty of optimizing three parameters simultaneously, the redundancies tend to make the final result dependent on the starting guess.

As an alternative to optimizing α, g and h simultaneously, it turned out to be beneficial to use different values of one of the parameters and run the minimization for the other two parameters at each of the selected values. The set of values that produce the minimum SSE is selected for further analysis. It is easy to show with a multiplicative transformation of the random variable that the parameter α is inversely related to the variance of the distribution [19]. Thus, if nothing is known about α, a rule of thumb for fitting an S-distribution to a data set is to compute the sample standard deviation and to select $1/s$ as a first value for α.

The strategy of fixing the value of α and leaving g and h to be fitted produces a set of quasi-equivalent S-distributions. These S-distributions are characterized by different triplets (α, g, h) of parameter values and are mathematically slightly different, yet have very similar residual errors (Figure 3) and essentially indistinguishable distributional shapes. While often noticeably different, the triplets are not scattered throughout the parameter space. For instance,

when plotted against α, the optimal values of the g and h parameter form well-defined regions. In particular, plots of g against $\log(\alpha)$ and against h are essentially straight lines. The characterization of these redundancies in parameters [25, 26] is very useful in practice, as we shall demonstrate below.

3. COMPUTING TRENDS IN AGE-DEPENDENT DATA USING S-DISTRIBUTIONS

3.1. General procedure

Given a set of age-dependent data, the goal is to estimate an S-distribution for each age group so that all S-distribution fits taken together provide an accurate description of the distributional trend as a function of age. Thus, it is not sufficient to determine one S-distribution per age class without consideration of neighbouring distributions. In addition to adequate individual fits, we desire a smooth variation in distributional shape throughout the entire age range. The following procedure yields such an overall fit:

1. For each age group, plot the sample median against age and fit the relationship with a polynomial or some other convenient, smooth function. Use this function for selecting a 'smoothed' median for each age group.
2. Compute the sample variance s^2 for each age group. Plot $1/s$, $2/s$ or $10/s$, or $1/s^2$ versus age and fit the relationship with a polynomial or some other convenient smooth function. Use this function for selecting a 'smoothed' α value for each age group. The combination of $\alpha = 1/s$ and a second-order polynomial has proven to be a good default. The particular choices are actually not as important as they might seem, as long as the trend with age is well represented by the approximating function.
3. Use the Mathematica© module described in Section 2.3 to determine for each age group the (g, h) pair that minimizes SSE. For this purpose, use the 'smoothed' α value from step 2 and initiate the numerical solution of the S-distribution at the 'smoothed' median from step 1 with $F(\text{smoothed median}) = 0.5$.
4. Plot the g values from step 3 versus age, and fit the relationship with a polynomial or another smooth function. This fit is used for selecting a 'smoothed' g values for each data group.
5. Using the 'smoothed' median, α and g values for each group, run another minimization just for h.
6. Plot the estimated h values from step 5 against age. Fit a polynomial in order to obtain 'smoothed' h values for any age.
7. Using the 'smoothed' parameters obtained in steps 1–6, compute the 'smoothed' S-distributions for each age group.

The results of this procedure depend to some degree on the choice of polynomials. High-order polynomials allow for more raggedness in trends, while polynomials of lower degree sometimes lead to a more pleasing representation of the trend. Clearly, the decision depends on the number of age classes, the purpose of the study, and on other information that may suggest that either a smooth or a rather ragged trend is more realistic. This ambiguity is not due to the methods proposed here, but germane to all smoothing algorithms of this type (for example, see discussion in Cole and Green [10]).

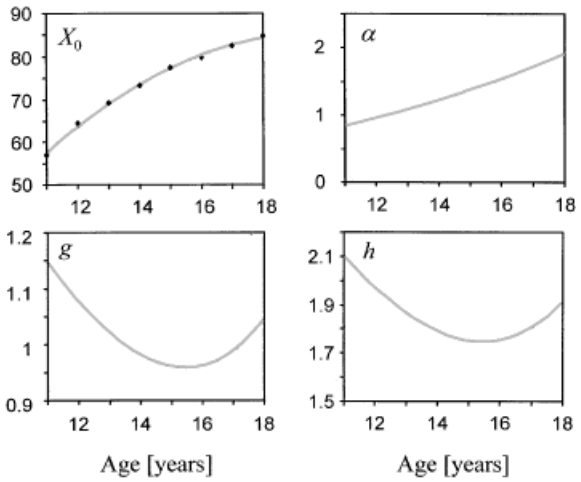

Figure 4. Trends in S-distribution parameters obtained after fitting the corresponding S-distributions to simulated data sets obtained from equation (3). These trends can be used to obtain – by interpolation – the corresponding S-distribution parameters for any age.

The results also depend on the sample sizes in each age group. If several age classes are grouped, the increased sample sizes are expected to yield more certain parameter estimates. At the same time, the loss in the number of classes leads to increased uncertainty in the estimation of trend parameters. Again, this is a general compromise in the field.

3.2. Application to simulated data

It is generally useful to test a new procedure first with simulated data. To this end, we define a scenario in which the conditional distribution for each age group is normal. Throughout the age range, these normal distributions exhibit artificial trends in mean and standard deviation of the form

$$\mu(\text{age}) = \frac{100 \times \text{age}^3}{1000 + \text{age}^3} \tag{3}$$

$$\sigma(\text{age}) = \frac{\mu(\text{age})}{10}$$

We generate samples for each age group from 11 to 18 years, with a sample size of 100 individuals each. Application of the procedure in Section 3.1 leads to a well-defined pattern in each of the S-distribution parameters (Figure 4), and to a corresponding trend in distributions.

Of course, the resulting representation of the actual trend depends on the sampling results. In this simulation the results are quite accurate when compared with the original distributions. To appreciate the performance of the method, we compare different percentile values. Percentiles are compared for an S-distribution by integrating the *inverted S-distribution equation* [19, 22]

$$\frac{\mathrm{d}X}{\mathrm{d}F} = \frac{1}{\alpha(F^g - F^h)} \quad X(0.5) = \text{median} \tag{4}$$

A comparison of percentiles obtained from the true distributions and from the fitted S-distributions reveals that the differences are small in all cases (Table III). A comparison of the S-distribution percentiles with percentiles of fitted normal distributions produces errors of a similarly small magnitude, attesting to the good fit of the S-distribution (results not shown).

In real applications, the characteristics of a trend may be much more complicated. Some examples include data on triceps skin folds in Gambian females from birth to age 50 years [10], weight and height growth of children [1,5], counts of CD4 lymphocytes versus age in non-HIV-1 infected children [9], and alkaline phosphatase measurements on girls from birth to 18 years [6]. Taking these situations as conceptual reference, we modelled complex trends to test the performance of the S-distribution. Three examples are shown in Figure 5. In the first example, data were generated from Weibull distributions with a quadratic trend in means (Figure 5(a)). In the second example (Figure 5(b)), a sinusoidal function was used as a trend for the mean. The underlying distributions were chosen to model strong heteroscedasticity. As a third example (Figure 5(c)), a fourth-order polynomial was used as a trend for the mean. As in the second example, strong heteroscedasticity was simulated. These examples demonstrate that the S-distribution method is capable of accurately estimating percentiles even in these more complex cases. As was observed with other methods (for example, Cole and Green [10]), the estimation at the extremes of the age range is somewhat less accurate than in the central values. Cole and Green discuss these *edge effects*, and we may add as a possible explanation that the first and last data sets are only constrained by one neighbouring distribution each, whereas all others are constrained by two neighbours.

3.3. Application to actual growth data

As a realistic illustration, we consider weights of 2631 girls from the Catalonia region in Spain [5]. The data span an age range from 5.5 to 17 years and are shown as bars in Figure 6; some summary statistics are given in Table II.

As a first quality assessment, one may fit each distribution in Figure 6 individually with an S-distribution. The results are not shown, but are actually slightly better than the superimposed densities in Figure 6, which represent S-distributions that are constrained by age trends in parameters (see below). For each individual (unconstrained) S-distribution, one easily computes all desired percentiles, as shown in Section 3.2, and these may be graphed along with the original data (Figure 7). Simply connecting the percentiles along age classes represents the extreme case in which the data *within* each class are smoothed by S-distributions, but the corresponding percentiles *among* classes are not smoothed (straight line connections of points) or minimally smoothed (for example, by spline). This representation constitutes the most detailed, least smooth characterization of the data that is based on S-distribution fitting.

The question now arises as to what degree the ups and downs in the percentile curves are true features of the entire population of Catalan girls of a given age, and to what degree they are idiosyncrasies of the specific samples. This is an unanswerable question, and much of the scientific discussion about percentile curves has dealt with the issue in general (for instance, see Cole and Green [10]). In fact, the smoothness of any trend line computed from the raw data or from something like the age class specific S-distribution percentiles is to some degree a matter of aesthetics and personal taste.

Two avenues can be pursued: an interpolation between percentiles of neighbouring age classes, or an estimation of trends in S-distribution parameters. The first option is easily executed, for

Table III. Comparison of percentile values for simulated data sets. Estimated percentiles correspond to S-distributions obtained from the procedure in Section 3.1. True values are percentiles computed from the original distributions.

Age	Estimated	True value	Relative difference
Percentile 3			
11	43.47	46.36	− 0.062
12	51.82	51.43	0.008
13	56.68	55.80	0.016
14	60.83	59.51	0.022
15	64.22	62.63	0.025
16	66.26	65.26	0.015
17	68.30	67.46	0.012
18	67.72	69.31	− 0.023
Percentile 10			
11	49.05	49.78	− 0.015
12	56.73	55.23	0.027
13	61.23	59.91	0.022
14	65.16	63.90	0.020
15	68.65	67.26	0.021
16	70.84	70.08	0.011
17	73.34	72.44	0.012
18	74.00	74.42	− 0.006
Percentile 25			
11	53.20	53.25	− 0.001
12	60.68	59.07	0.027
13	65.19	64.09	0.017
14	69.11	68.35	0.011
15	72.85	71.94	0.013
16	75.11	74.96	0.002
17	77.82	77.48	0.004
18	79.37	79.61	− 0.003
Percentile 50			
11	56.94	57.10	− 0.003
12	64.46	63.34	0.018
13	69.24	68.72	0.008
14	73.29	73.29	0.000
15	77.47	77.14	0.004
16	79.73	80.38	− 0.008
17	82.44	83.09	− 0.008
18	84.79	85.36	− 0.007
Percentile 75			
11	60.43	60.95	− 0.008
12	68.16	67.62	0.008
13	73.39	73.36	0.000
14	77.69	78.23	− 0.007
15	82.46	82.35	0.001
16	84.64	85.80	− 0.013
17	87.20	88.69	− 0.017
18	90.30	91.12	− 0.009

Table III. (Continued)

Age	Estimated	True value	Relative difference
Percentile 99			
11	71.07	70.38	− 0.010
12	79.74	78.08	0.021
13	86.88	84.71	0.026
14	92.24	90.34	0.021
15	99.34	95.09	0.045
16	101.08	99.08	0.020
17	102.68	102.42	0.003
18	108.16	105.22	0.028

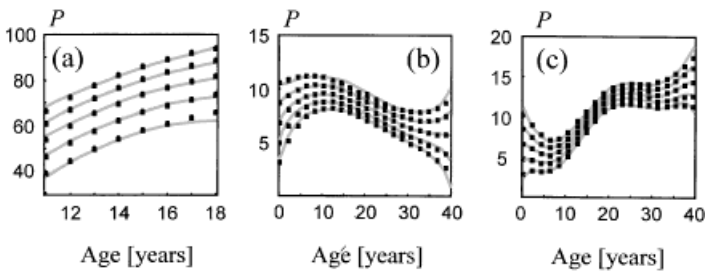

Figure 5. Estimated percentile (P) curves for simulated data. Data are generated upon defining age-dependent trends in median and shape. Dots indicate percentiles computed from the original distributions. Curves exhibit trends in percentiles estimated with the S-distribution method. In each case, the underlying age-related distributions and the corresponding trends in mean and variance were inspired by published cases: (*a*) Weibull distributions with quadratic trend in the mean; (*b*) a sinusoidal function is used as a trend for the mean – the process is strongly heteroscedastic; (*c*) a fourth-order polynomial is used as a trend for the mean – the process is strongly heteroscedastic.

instance, in Excel©. Figure 7 shows such interpolations with sixth-order polynomials. Indeed, these interpolations are very similar to the curves resulting from the interpolation method of Puente *et al.* [5] (comparison not explicitly shown, however compare Figures 7 and 8, right panel).

The second option follows the procedure outlined in Section 3.1. It begins with the selection of a polynomial capturing the age-dependent trend in medians. Evaluating several alternatives, we decided to use a fifth-order polynomial. A lower-order polynomial would lead to smoother results but carry the risk of ignoring true local changes in trend, while a polynomial of very high order could become unrealistically ragged. The polynomial chosen for the age-dependent medians has the form

$$\text{Median(age)} = -231.1 + 129.9 \times \text{age} - 26.09 \times \text{age}^2 + 2.532 \times \text{age}^3 - 0.1162 \times \text{age}^4 + 0.002027 \times \text{age}^5 \quad (5)$$

Further executing the procedure proposed in Section 3.1, the following polynomials were selected for capturing the changes in the remaining S-system parameters:

$$\alpha(\text{age}) = 1.8080 - 0.2031 \times \text{age} + 0.0071 \times \text{age}^2$$

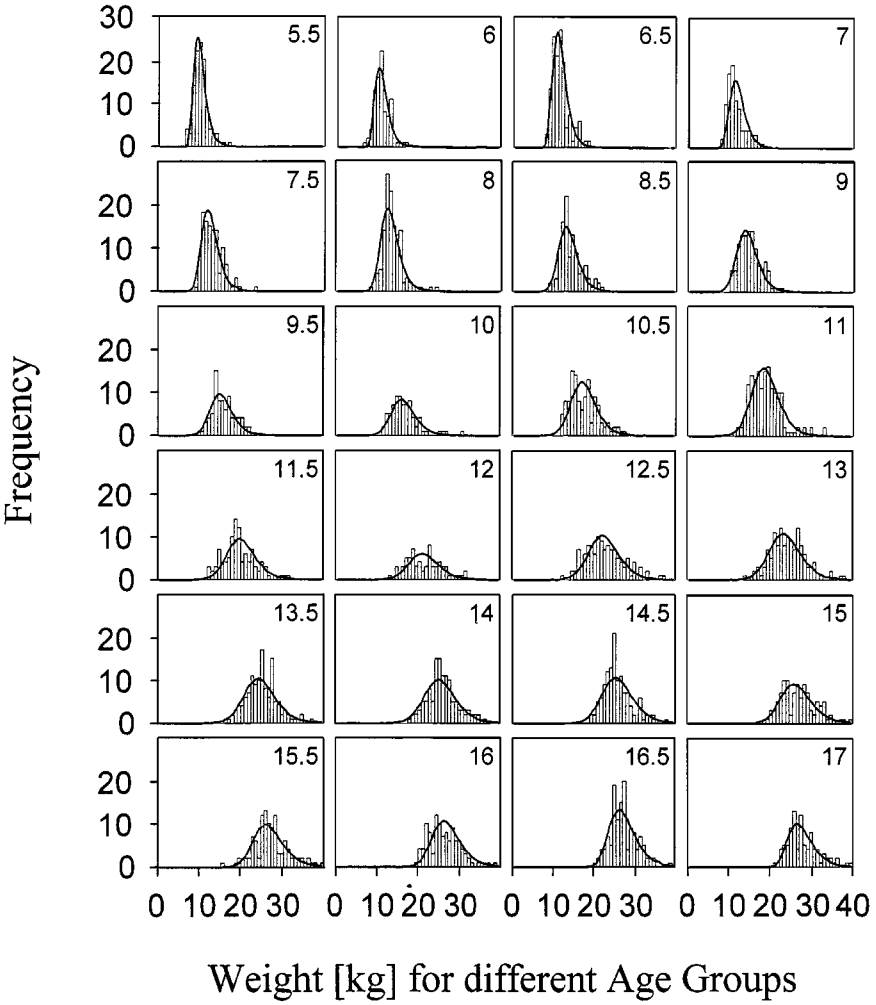

Figure 6. Histograms of weights of girls between ages 5.5 and 17 years (top right corner of each panel), overlaid with fitted S-distributions. Descriptive statistics for the data are presented in Table II. Data were redrawn from Puente *et al.* [5].

$$g(\text{age}) = 0.6236 + 0.0077 \times \text{age}^2 - 0.00041 \times \text{age}^3$$
$$h(\text{age}) = 1.0382 + 0.0088 \times \text{age}^2 - 0.00046 \times \text{age}^3 \quad (6)$$

In the cases of $\alpha(\text{age})$ and $g(\text{age})$, the polynomials are used as intermediate refinement steps for selecting appropriate parameter values for the final step of obtaining h. The fitted h values show a definite trend that is accurately modelled by $h(\text{age})$ in equation (6) ($r^2 = 0.997$).

As the α, g and h parameters determine the spread and shape of the underlying S-distribution, it is clear that higher-order polynomials in these parameters over age would lead to distributions and associated percentiles that would more closely resemble the unconstrained percentiles in

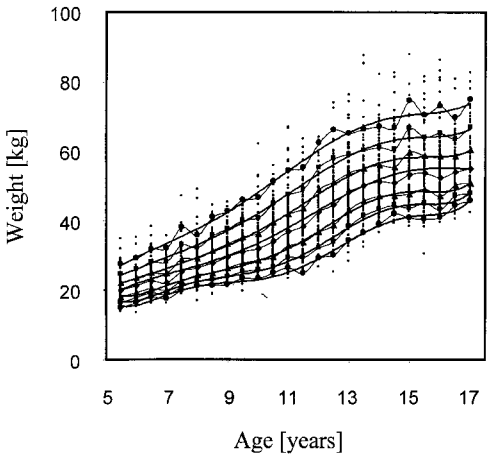

Figure 7. Comparison of raw data, percentiles of S-distributions that were fitted for each age class without regard of constraints between classes, and interpolating sixth-order polynomial trend lines. The centre trend line was computed from the raw data, while all other trend lines were computed from S-distribution percentiles. Data redrawn from Puente *et al.* [5]; one outlying data point was considered in the computations but not graphed.

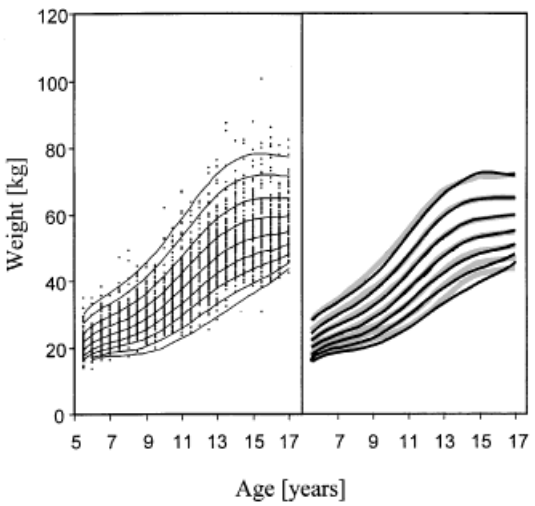

Figure 8. Percentile curves for the weights of Catalan girls, ages 5.5 to 17 years [5]. Left panel: comparison of observed distributions of individual weights with percentiles curves obtained from the estimated trend in S-distributions. From bottom to top, the percentile curves are 1, 3, 5, 25, 50, 75, 95, 97 and 99 per cent. Right panel: comparison of two methods for fitting percentile curves. Black curves represent percentiles of S-distributions (as in left panel). Grey lines connect percentiles estimated by Puente *et al.* [5]. From bottom to top, the percentile values are 3, 5, 25, 50, 75, 95 and 97 per cent.

Figure 7 (symbols). Thus, the chosen functions in equations (5) and (6) are compromises, subject to discussion.

Execution of the smoothing method reveals a clear trend in the age-related distributions, which can be seen in the PDFs themselves (see Figure 6) or in the percentile curves (Figure 8). The left

panel of Figure 8 shows how the smoothed percentiles compared with the individual data. It is noted that the percentile curves are not strictly parallel, which is an indication of the non-normality and the heteroscedasticity in the data. The 99 per cent percentile curve shows a maximum around age 15, which, interestingly, is very similar to the maximum observed in American girls [10]. The right panel of Figure 8 compares the results with those obtained by Puente *et al.* [5] using interpolation techniques. While the central percentile curves are rather similar, differences clearly exist in the lower percentiles of older age groups. Inspection of the data in the left panel explains this difference. For instance, for ages 14 to 15, the reported minimum weights are much higher than for the neighbouring classes. As a simple interpolation, the method used by Puente *et al.* [5] reflects this phenomenon, while the S-distribution method with the relatively low-order polynomials chosen above smoothes it out. Whether the phenomenon is real or a coincidental sampling effect cannot be decided from the data. A similar phenomenon creates slight differences in the upper percentiles around age 9. The decision between smoother or more ragged percentile curves is a well-known conundrum (see Cole and Green [10]), and if a more detailed representation were preferred over the smoother trend lines shown here, one would simply use higher-order polynomials or other functions to capture the development of the S-system parameters over the observed range of age classes (see above).

The age-dependent parameter values allow us to estimate weights of girls of any age within the given range, along with any percentiles of interest. For example, the results suggest that the S-distribution of weights of 12.5 year old girls has the parameter values $F(45.2) = 0.5$, $\alpha = 0.379$, $g = 1.027$ and $h = 1.510$. Percentiles of interest are obtained by integrating this S-distribution. For instance, the percentile corresponding to a 12.5-year-old girl weighing 40.2 kg is 24.8. The same weight percentile for 12-year-old girls is considerably higher at 35.8.

One of the important uses of age-specific reference intervals is the characterization of extreme values. Continuing with the same example, a 12.5-year-old girl weighing 35 kg corresponds to a percentile of 8.1, while a weight of 31 kg corresponds to a percentile of 2.7. Very low or very high percentiles may be used as guidelines for thresholds outside which further screening for pathologies might be recommended.

DISCUSSION

The computation of age-specific percentile curves of a biological marker provides an important tool for a first screening of potentially pathological situations of public health concern. We have presented a new parametric method based on S-distributions for computing such curves. The approach estimates conditional distributions as functions of age by employing a fitting procedure that reveals age trends in the S-distribution parameters. Simulation studies suggest that the methodology yields accurate results and that it can deal with complex trends. An analysis of actual data furthermore demonstrates that percentile curves computed with the S-distribution method are close to those estimated by other techniques, even though the similarity of results depends on the chosen degree of smoothness. If the data are fitted within each class and the resulting S-distribution percentiles are locally interpolated with splines, the results are essentially equivalent with those of Puente *et al.* [5]. Additional smoothing among classes gradually eliminates the raggedness of the percentile curves, but risks missing true deviations from expected smoothness.

The proposed method can be applied to data distributions with vastly differing shapes. The theory behind S-distributions and a growing body of experience suggest that parameter

combinations exist that reproduce the known, relevant types of unimodal distributions, as well as distributions that are difficult to model with any traditional distribution, such as those extremely skewed to the left.

Although the S-distribution method was introduced in the context of age-related percentile curves, the same techniques can be applied to conditional distributions that are functions of other covariates [22, 23]. In its present form, the method requires groups represented by a single covariate. In many applications this appears an acceptable strategy that yields useful results. None the less, if needed, it seems plausible to extend the method to more complex groupings by using a strategy similar to the one discussed by Healy *et al.* [3].

While consistency with other methods is a good indication of quality, any new method should show advantages beyond the current standard. We see these advantages in two differentiating aspects. First, the proposed method does not require the original data to be transformed. While transformations, such as the one proposed by Box and Cox [11] and used in the *LMS* method [10], may in many cases be straightforward, there is no guarantee that the resulting distributions are sufficiently normal. As Cole and co-workers demonstrated, the *LMS* method captures trends in distributions with moderate changes in skewness well. It remains to be seen whether the *LMS* method allows for the same degree of shape flexibility and possibly trend reversal as methods based on S-distributions, namely the method proposed here or a more complicated S-distribution method presented elsewhere [27].

Secondly, there is no doubt that the method proposed here naturally emphasizes the *commonality* [10] among distributions in different age classes. In addition to guaranteeing commonality between neighbouring classes, the S-distributions for *all* age classes are intimately related, since they all fall into the same mathematical structure. The parameters of the different S-distributions are constrained through some smooth functions. These functions can be chosen almost arbitrarily, as long as they represent the individual parameter trends well. This choice obviously introduced some arbitrariness, which however is typical and necessary for all smoothing techniques in that it allows the researcher to pre-set the level of smoothness or raggedness.

ACKNOWLEDGEMENTS

This work was partially supported by a grant CICYT PM099-99 and a grant from La Paeria. Weight data on girls were obtained in a cross-sectional study in the region of Catalonia (Spain) and were kindly provided by the Unitat Clinico-Epidemiologica de la Ciutat Saniraria i Universitaria de Bellvitge. We thank Marta Fajó for her collaboration in the early stages of this research.

REFERENCES

1. Cole TJ, Freeman JV, Preece MA. British 1990 growth reference percentiles for weight, height, body mass index and head circumference fitted by maximum penalized likelihood. *Statistics in Medicine* 1998; **17**:407–429.
2. Bonellie SR, Raab GM. A comparison of different approaches for fitting percentile curves to birth weight data. *Statistics in Medicine* 1996; **15**:2657–2667.
3. Healy MJ, Rasbash J, Yang M. Distribution-free estimation of age-related percentiles. *Annals of Human Biology* 1988; **15**:17–22.
4. Pan HQ, Goldstein H, Yang Q. Non-parametric estimation of age-related percentiles over wide age ranges. *Annals of Human Biology* 1990; **17**:475–481.
5. Puente ML de la, Canela J, Alvarez J, Frenández-Goula ME, Lara N de, Martí C, Jiménez A, Rue M, Coll JJ, Barredo M, Callís L, Vicens-Calvet E, Salleras L. Estàndards transversals de creixement de la població infantil i adolescent de Catalunya (1986–87). *Butlleti de la Societat Catalana de Pediatria* 1993; **53**:251–256.

6. Tango T. Estimation of age-specific reference ranges via smoother AVAS. *Statistics in Medicine* 1998; **17**:1231–1243.
7. Royston P, Matthews JNS. Estimation of reference ranges from normal samples. *Statistics in Medicine* 1991; **10**:691–695.
8. Royston P. Constructing time-specific reference ranges. *Statistics in Medicine* 1991; **10**:675–690.
9. Wright EM, Royston P. Simplified estimation of age-related reference intervals for skewed data. *Statistics in Medicine* 1997; **16**:2785–2803.
10. Cole TJ, Green PJ. Smoothing reference percentile curves: the LMS method and penalized likelihood. *Statistics in Medicine* 1992; **11**:1305–1319.
11. Box GEP, Cox DR. An analysis of transformations. *Journal of the Royal Statistical Society*, *Series B* 1964; **26**:211–252.
12. Mohler CL, Marks PL, Sprugel DG. Stand structure and allometry of trees during self-thinning of pure stands. *Journal of Ecology* 1978; **66**:599–614.
13. Ford ED. Competition and stand structure in some even-aged plant monocultures. *Journal of Ecology* 1975; **63**:311–333.
14. Gates DJ, McMurtrie R, Borough CJ. Skewness reversal of distribution of stem diameter in plantations of Pinus radiata. *Australian Forest Research* 1983; **13**:267–270.
15. Pearson K. Contributions to the mathematical theory of evolution. II. Skew variations in homogeneous material. *Philosophical Transactions of the Royal Society of London*, *Series A* 1895; **186**:343–414.
16. Johnson NL, Kotz S. *Continuous Univariate Distributions*—1. Houghton Mifflin: Boston, MA, 1970.
17. Savageau MA. A suprasystem of probability distributions. *Biometrical Journal* 1982; **24**:323–330.
18. Voit EO, Rust PF. Tutorial: S-system analysis of continuous univariate probability distributions. *Journal of Statistics and Computer Simulation* 1992; **42**:187–249.
19. Voit EO. The S-distribution: A tool for approximation and classification of univariate, unimodal probability distributions. *Biometrical Journal* 1992; **34**:855–878.
20. Voit EO, Yu S. The S-distribution, approximation of discrete distributions. *Biometrical Journal* 1994; **36**:205–219.
21. Yu S, Voit EO. A simple, flexible failure model. *Biometrical Journal* 1995; **37**:595–609.
22. Voit EO. Balthis WL, Holser RA. Hierarchical Monte-Carlo modeling with S-distributions: concepts and illustrative analysis of mercury contamination in king mackerel. *Environmental International* 1995; **21**:627–635.
23. Balthis WL, Voit EO, Meaburn GM. Setting prediction limits for mercury concentrations in fish having high bioaccumulation potential. *Environmetrics* 1996; **7**:429–439.
24. Voit EO, Savageau MA. Analytical solutions to a generalized growth equation. *Journal of Mathemical Analysis and Applications* 1984; **103**:380–386.
25. Sands PJ, Voit EO. Flux-based estimation of parameters in S-systems. *Ecological Modelling* 1996; **93**:75–88.
26. Berg PH, Voit EO, White R. A pharmacodynamic model for the action of the antibiotic Imipenem on Pseudomonas in vitro. *Bulletin of Mathematical Biology* 1996; **58**:923–938.
27. Voit EO. Dynamic trends in distributions. *Biometrical Journal* 1996; **38**:587–603.

