The Construction of Standards for Measurements Subject to Growth

By H. Goldstein[1]

ABSTRACT

A general method for constructing population centiles of a measurement subject to growth, is proposed. The method requires the fitting of regression lines of the mean value of the measurement on age, within a series of narrow age ranges. The deviations from these lines are then used to estimate the distances of the centiles from these lines.

The use of standards for growth attained at different ages is now well established (Tanner, Whitehouse, and Takaishi, 1965). Relatively little attention however has been given to the statistical problems of estimating the population centiles.

This paper is concerned only with the construction of "distance" standards, that is population standards for a measurement at a series of ages. The usual method of constructing such standards is to estimate centiles at a number of ages, for example at every year of age, and then to pass smooth curves through these points so as to cover the whole age range.

If we consider a particular age, say 7.0 years, all the children will not in general be measured at exactly 7.0 years; for example we may have available a sample of children aged between 6.5 and 7.5 years. Healy (1962) points out that these measurements cannot be used directly to estimate, say, the standard deviation at age 7.0, since the fact that the mean value changes over this age range implies that the sample standard deviation will be larger than the "instantaneous" figure which is required. Healy derives a correction to be applied to the sample variance in order to give an unbiased estimate of the instantaneous variance. In his derivation, Healy assumes that the mean and variance increase linearly with age over the age range and that all ages in the range are equally represented in the sample. In addition to these two

[1] Department of Growth and Development Institute of Child Health, 30 Guilford Street, London, W.C. 1. Present address: National Children's Bureau, 1 Fitzroy Square, London, W1P 5AH.

Human Biology, May, 1972, Vol. 44, No. 3, pp. 255-261

requirements the method can be used only when the measurement (or a transformation of it) is assumed to be normally distributed. In some applications the first two requirements may not be satisfied. In a longitudinal study for example, measurements are usually clustered around specific ages, often birthdays, and the assumption of a uniform age distribution cannot therefore be made. If the age range is very broad, a non-uniform distribution may also occur, and moreover the relationship between the mean value of the measurement and age may be markedly non-linear. The present paper extends Healy's results by dropping the first two requirements, and will also suggest ways in which non-normal data might be treated.

Method

The underlying idea is to allow for the change in mean value over the age range by fitting a trend line to the sample data, and then to use the deviations of the sample points from the fitted line to estimate the centiles.

Consider a measurement y taken at age t. For a given age range we have

$$E(y) = f(t) \tag{1}$$

In the following discussion we shall assume that $f(t)$ is a linear function of t, although the method applies equally well to non-linear functions. For the ith sample measurement therefore we have

$$y_{ij} = \alpha + \beta t_i + \epsilon_{ij} \tag{2}$$

where the α, β are to be estimated from the sample, and the y_{ij}, t_i are known. The deviations from the line, the ϵ_{ij}, will be referred to as 'residuals'. In general the variance of the ϵ_{ij} will depend on t_i and we may write

$$\sigma^2(\epsilon_{ij}) = g(t_i) \tag{3}$$

we also assume

$$E(\epsilon_{ij}) = 0 \qquad \text{For all i,j}$$

We will also assume in what follows that

$$\sigma^2(\epsilon_{ij}) = \sigma_i^2 = (1 + \gamma t_i)\sigma^2 \tag{4}$$

although as above, the method does not necessarily require this linearity assumption.

Goldstein, H., *The Construction of Standards for Measurements Subject to Growth* , Human Biology, 44:2 (1972:May) p.255

Least squares estimates of α,β in (2) are given by minimising, with respect to α,β the expression

$$\sum_{i,j} w_i(y_{ij} - \alpha - \beta t_i)^2 \qquad (5)$$

where

$$w_i = (1 + \gamma t_i)^{-1}$$

We shall assume that the parameter γ and therefore the w_i are known. If not, they can be separately estimated from the sample (see discussion). The solution of (5) when these parameters are unknown is difficult.

For estimators of α, β we have

$$\hat{\beta} = \sum_{i,j} w_i(t_i - \bar{t})(y_{ij} - \bar{y}) \;/\; \sum_{i,j} w_i(t_i - \bar{t})$$

$$\hat{\alpha} = (\bar{y} - \hat{\beta}\bar{t}) \;/\; \sum_{i,j} w_i$$

where

$$t = \sum_{i,j} w_i t_i \;/\; \sum_{i,j} w_i$$

$$y = \sum_{i,j} w_i y_{ij} \;/\; \sum_{i,j} w_i$$

If the sample size is large the weighted residuals from the fitted line

$$r_{ij} = (y_{ij} - \alpha - \beta t_i) w_i^{1/2} \qquad (6)$$

may be used to provide the estimates of the population centiles about the line. An unbiased estimate of σ^2 is given by

$$\sum_{i,j} r_{ij}^2/(n - 2) \qquad (7)$$

where n is the sample size, and if the distribution of the r_{ij} is assumed to be normal, this may be used to construct centiles in the usual way using (4) to calculate the variance at t_i, for example at the centrc of the age interval.

If the normality assumption is not made we may still make direct estimates of population centiles if we assume that the r_{ij} have the same distribution at all points on the time scale. Since the first two moments do not depend on t, this involves assuming that the higher order moments of the distribution also do not depend on t. In any given application the plausibility of this assumption may be studied by calculating

say, the third and fourth moments of the r_{ij} for different values of t. If the assumption is accepted (see discussion) we may then proceed by ordering the r_{ij} in increasing order of magnitude and estimating the centile values at the corresponding points in this ordering. A practical technique for doing this is to plot the points on normal probability paper, so that at the appropriate centile values a curve can be fitted by eye to surrounding points to obtain a smoothed estimate of the centile value. This is then multiplied by $(1+\gamma t_i)^{1/2}$ and added to the value predicted by the regression line at age t_i, to give the centile estimate of the measurement at this age.

Discussion

In the above derivation we have assumed that an estimate of γ is independently available. This, however, will only be true for certain measurements in certain populations. In other cases γ will have to be estimated from the sample. If the sample consists of a large number of measurements at each of a number of discrete ages, then the variance can be estimated at each age and an estimate of γ readily obtained. Usually, however, we do not have such a sample available, and in this case the following procedure may be used. The age range is divided into a number of smaller intervals each containing approximately the same number of individuals. The intervals should be narrow enough to be able to assume approximately the same variance at each age within the interval, but wide enough to obtain a large enough sample to give a reliable estimate of that variance. A preliminary scatterplot of the data could be a useful aid in deciding which divisions to make. Within each division an unweighted linear regression line of the measurement on age is fitted to the data, and the residual variance estimated. This variance is then taken to be the instantaneous variance at the centre of the age interval. If the scatterplot suggests a non linear relationship then a higher order polynomial should be fitted. Finally an unweighted linear regression of these variances on age is fitted for the set of intervals, and γ is estimated as b/a where b is the slope and a is the intercept of the regression line.

When examining the third and fourth moments of r_{ij} the age range should similarly be divided into intervals and these moments estimated within each interval, and examined for any dependence on age.

It may happen that we cannot choose narrow enough intervals such that all the moment estimates are approximately constant within each interval. In particular a dependence of the moments on age may

be distorted by changes with age which differentially affect the estimates in each interval. In this case we should make another choice of interval and recompute the above statistics, to check whether this changes the relationships with age. A suitable other choice would be one where the intervals were the same width (subject to the numbers within the intervals not being too small).

Finally, there is the problem of the width of the age range to be used. If the data extends over several years we might treat this either as a single age range, or split it into several, for example, yearly age ranges and carry out the calculations separately for each one, combining the separate estimates by smoothing them over the whole range (Tanner et al., 1966). The decision on how many age ranges to use will depend on how well the assumptions of the above method can be satisfied. In general it would seem desirable to keep the age ranges as small as possible, consistent with keeping sufficient sample numbers within the range.

Numerical Illustration

The data used in this example are a random subsample of 655 girls aged 7.0-7.5 years, from the National Child Development Study (Goldstein, 1971). Measurements of height were made to the nearest inch by untrained measurers. To carry out a weighted analysis we need to know the value of γ in (4) and this has been estimated from the standards of Tanner et al. (1966) to be 0.10 cm^2/yr for this age range, with age t_i measured from 7.0 years.

The equation of the weighted linear regression of height (y) on age (x) is: $y = 118.0 + 6.25x$ and with standard error of the regression coefficient = 1.87cm. The estimate of σ^2 which is also the residual variance at age 7.0 yr is 39.7 cm^2. The corresponding estimate from an unweighted analysis is 41.0 cm^2. Since height is known to be approximately normally distributed the estimate of the residual standard deviation may be used to construct centiles. At 7.0 yr the weighted estimate is 6.3 cm and thus the mean height predicted by the regression line at age 7.0 yr is 118.0 cm, and so for example, the 10th and 90th centiles are 118.0 ± 8.1 cm, or 109.9 cm and 126.1 cm.

Figure 1 shows a normal probability plot of the weighted residuals. The estimate of the 10th centile (110.0 cm) is approximately the same as that based on the standard deviation. At the lower end of the distribution however, the plot deviates from the straight line found in the middle of the distribution, and the smoothed estimate of the 3rd centile

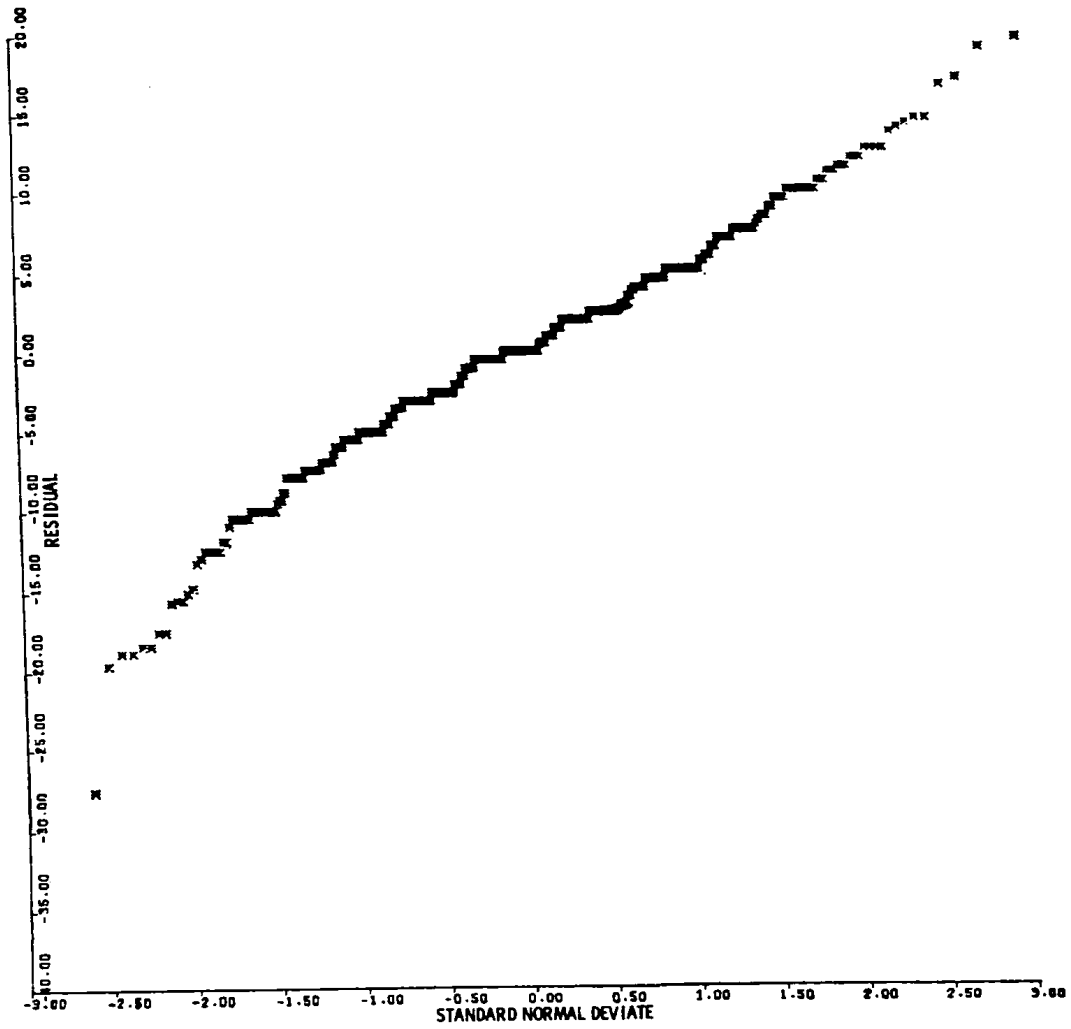

FIG. 1. Weighted Residuals from Regression Line of Height on Age. Ages 7.0-7.5.

is 105.2 cm which is appreciably different from the estimate of 106.1 cm based on the standard deviation. The 'step-like' appearance of the plot is due to the measurements having been taken to the nearest inch.

ACKNOWLEDGMENTS

I express thanks to the directors and steering committee of the National Child Development Study for permission to publish these data. My thanks are also due to Mr. M. J. R. Healy for helpful advice and criticism. This work was partly supported by a grant from the Nuffield Foundation to the Department of Growth and Development at the Institute of Child Health.

Literature Cited

Goldstein, H. 1971 Factors influencing the height of seven year old children: results from the National Child Development Study. Human Biol. **43**: 92-111.

Healy, M. J. R. 1962 The effect of age-grouping on the distribution of a measurement affected by growth. Am. J. Phys, Anthrop. **20**: 49-50.

Tanner, J. M., R. H. Whitehouse and M. Takaishi 1966 Standards from birth to maturity for height, weight, height velocity, and weight velocity: British Children, 1965. Arch. Dis. Child. **41**: 454-471, 613-635.