Modelling the Distribution of Anthropometrics: Gaussian Distributions, LMS Distributions, and Probability Plots

Jonathan Thornburg[1,2] and Virginia J. Vitzthum[2,3,4]

[1] Center for Spacetime Symmetries, Indiana University, Bloomington, Indiana, USA [2] BKIS Orchards, Thetis Island, BC, Canada [3] Dept. of Medicine, University of British Columbia, Vancouver BC Canada [4] Dept. of Anthropology, Indiana University, Bloomington, Indiana, USA

Introduction

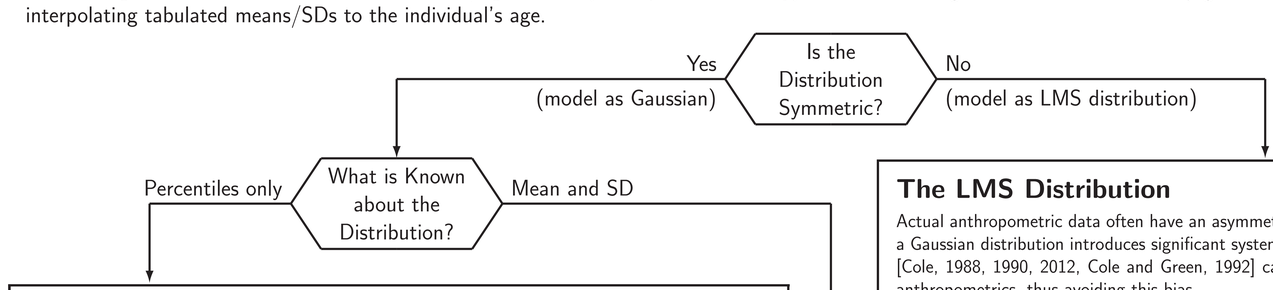
In anthropologists' studies of growth and development, particularly in non-industrialized populations, sample sizes are often small and span a range of ages. These attributes can hamper analyses and hypothesis testing, and make comparisons to other populations difficult. z-scores are a commonly used statistic to mitigate these challenges. A z-score is the value of an individual's anthropometric (y) expressed in units of the standard deviation (SD) of the anthropometric for a suitable sex- and age-specific reference sample. That is, $z = (y-\mu)/\sigma$ where μ and σ are the mean and SD of the reference sample, respectively. Depending on the research question, commonly used reference samples (e.g., WHO, CDC) are not necessarily suitable for all populations. Therefore, there are increasing efforts to construct population-specific growth references.

If a growth reference provides the mean and SD for each age/sex bin, it's easy to compute the z-score corresponding to any individual's measurement by assuming a Gaussian (normal) distribution. z-scores may be computed by either using the mean/SD for the individual's sex/age bin, or (for improved accuracy) interpolating tabulated means/SDs to the individual's age.

However, if the anthropometric has an asymmetric (skewed) distribution (as do weight, BMI, and many skinfolds), this approach results in systematically biased z-scores. Cole's LMS distribution can accurately represent the distributions of these and many other anthropometrics, avoiding this bias.

But sometimes only percentiles are provided for each age/sex bin. We describe how to: (a) determine the mean/SD of the Gaussian distribution, or the coefficients of the LMS distribution, that best fit the published percentiles; (b) visually assess the quality of such a fit; and (c) extrapolate the distribution beyond the range of the published percentiles and visually assess the quality of such an extrapolation.

We describe doing this fitting with common open-source software (Gnuplot, R, or SciPy (Python)), or with Microsoft Excel™. The fitted coefficients can then be used to compute z-scores. We also describe how to extrapolate parameters (and thus compute z scores) for an individual who is outside the tabulated age range, and we present a graphical assessment of any given extrapolation's quality.



Fitting a Gaussian to Percentiles

If we know a set of percentiles of a Gaussian distribution (say, data Y_i are the P_i% percentile for $i = 1, \dots, n$) we can find the Gaussian distribution's mean μ and SD σ using a **Gaussian probability plot**. This is a scatterplot where each known percentile is plotted at the (x, y) position

$$x_i = \texttt{Gaussian\_z\_of\_prob}(P_i/100)$$
$$y_i = Y_i$$

where $\texttt{Gaussian\_z\_of\_prob}(P_i/100)$ is the z-score corresponding to the percentile P_i%. (See the endnotes for more details on the `Gaussian_z_of_prob` function.)

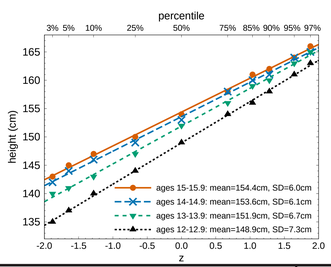
If the data are indeed percentiles of a Gaussian distribution, then the points will all be on a straight line with slope σ and y-intercept μ. In other words, by *fitting a straight line* to the points on a probability plot, we can determine the Gaussian's mean (the y-intercept of the fitted line) and SD (the slope of the fitted line).

Example

Baya Botti et al. [2009] give the following percentiles of girls' height (cm) in a nationally representative sample of Bolivian adolescents:

age group	3%	5%	10%	25%	50%	75%	85%	90%	95%	97%
12–12.9	135	137	140	144	149	154	156	158	161	163
13–13.9	140	141	143	147	152	156	159	160	163	165
14–14.9	142	144	146	149	153	158	160	161	164	165
15–15.9	143	145	147	150	154	158	161	162	164	166

The figure at right shows a Gaussian probability plot of these data, with a separate straight line fitted for each age bin. The points for each age bin are indeed very close to being on the bin's fitted straight line; this demonstrates that the distribution of girls' heights in this population is very close to Gaussian for each age bin.



The LMS Distribution

Actual anthropometric data often have an asymmetric (skewed) distribution, so modelling them with a Gaussian distribution introduces significant systematic errors in the z-scores. The LMS distribution [Cole, 1988, 1990, 2012, Cole and Green, 1992] can accurately represent the distributions of many anthropometrics, thus avoiding this bias.

The LMS distribution is defined by transforming a Gaussian distribution, and has 3 parameters:

- L specifies the skewedness of the LMS distribution:
  - for $L = 1$ the LMS distribution is equal to a Gaussian;
  - for $L < 1$ the LMS distribution has a long right tail;
  - for $L > 1$ the LMS distribution has a long left tail.
- M is the median of the LMS distribution (M is also the mean of the original Gaussian distribution).
- S is the CV (coefficient of variation, i.e., SD/mean) of the original Gaussian distribution, so that MS is the SD of the original Gaussian distribution. For L not too different from 1, S is approximately the CV of the LMS distribution and MS is approximately the SD of the LMS distribution.

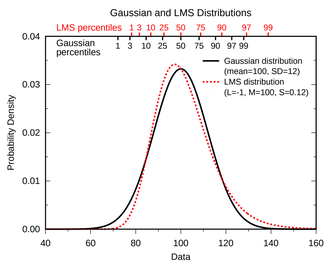
LMS data y are related to a standard Gaussian z-score (a Gaussian with mean 0 and SD 1) by

$$y_{\text{LMS}}(z) = M(LSz + 1)^{1/L}, \qquad (1)$$

and, correspondingly, the P% percentile of an LMS distribution is given by

$$\texttt{LMS\_z\_of\_prob}(P/100) = y_{\text{LMS}}\Big(\texttt{Gaussian\_z\_of\_prob}(P/100)\Big).$$

The figure at right shows a comparison of example Gaussian (black) and LMS distributions (red). Notice that even though the distributions appear quite similar, their percentiles are substantially different, particularly near the extremes (tails) of the distribution.



Fitting an LMS Distribution to Percentiles

If we know a set of percentiles of an LMS distribution (say, data Y_i are the P_i% percentile for $i = 1, \dots, n$), how can we find the LMS distribution's L, M, and S parameters? That is, how can we fit an LMS distribution to a set of known percentiles? Cole [1990] and Cole and Green [1992] describe methods to do this, but they are mathematically complicated.

Here we present a simpler fitting method which leverages existing widely-available software packages:

1. Plot a Gaussian probability plot of the percentiles. We emphasize that here the z-score corresponding to each percentile P_i% is to be computed using the Gaussian `Gaussian_z_of_prob`$(P_i/100)$ function. Typically the points will fall along a smooth curve, but not on a straight line.
2. Estimate the mean μ and SD σ of a Gaussian which approximately fits the percentiles, by fitting a straight line to the percentiles on the Gaussian probability plot.
3. Nonlinearly fit an LMS distribution to the percentiles on the Gaussian probability plot, using the fitted Gaussian as an initial guess.

   More precisely, nonlinearly fit the model (1) to the (z, y) data points

   $$z_i = \texttt{Gaussian\_z\_of\_prob}(P_i/100)$$
   $$y_i = Y_i,$$

   using the initial guess $L = 1$, $M = \mu$, $S = \sigma/\mu$. (See the endnotes for more details on nonlinear fitting.)
4. Plot the fitted LMS distribution (i.e., plot the model (1)) on the Gaussian probability plot and verify that the points closely fit the model.

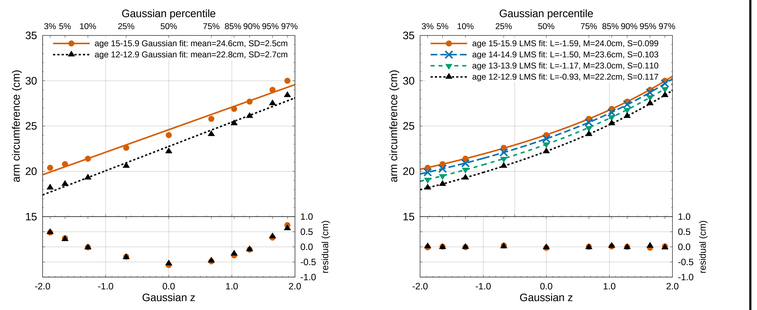
Example

Baya Botti et al. [2009] give the following percentiles of girls' arm circumference (cm) in a nationally representative sample of Bolivian adolescents:

age group	3%	5%	10%	25%	50%	75%	85%	90%	95%	97%
12–12.9	18.2	18.6	19.3	20.6	22.2	24.1	25.3	26.1	27.5	28.4
13–13.9	19.1	19.5	20.2	21.4	23.0	24.8	26.0	26.8	28.2	29.1
14–14.9	19.9	20.3	20.9	22.1	23.6	25.4	26.5	27.4	28.7	29.7
15–15.9	20.4	20.8	21.4	22.6	24.0	25.8	26.9	27.7	29.0	30.0

The left figure below shows the arm-circumference percentiles and the best-fitting Gaussian for the 12–12.9 and 15–15.9 years age bins (step 2 above). Notice that while the Gaussian is a reasonable approximation, the actual percentiles differ systematically from the Gaussian percentiles, with residuals as high as $\frac{3}{4}$ cm for some parts of the distribution.

The right figure below shows the result of fitting the LMS model (1) to each age bin's percentiles, (step 4 above). Each age bin's model fits very well, i.e., the distribution of girls' arm circumference in this population is very close to an LMS distribution. The residuals (which are plotted at the same scale as the Gaussian-fit plot) are now $< \frac{1}{2}$ mm everywhere in the distribution, with no systematic pattern.

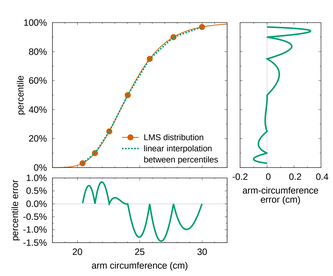


Calculating z-scores and/or Percentiles

Once the parameters of a distribution (mean and SD for a Gaussian, or L, M, and S for an LMS distribution) are known, the z and/or percentile corresponding to any given anthropometric can be calculated. This is sometimes done by linearly interpolating between tabulated percentiles, but it's preferable to use the `Gaussian_z_of_prob` or `LMS_z_of_prob` function, for two reasons.

First, interpolation is only possible for anthropometrics within the range of the tabulated percentiles. For a sample size of 12 or more, there's a > 50% chance of at least one individual falling outside the 3% to 97% percentile range.

Second, the interpolation introduces additional errors. These are generally small if the tabulated percentiles are relatively finely spaced, but may be substantial if the known percentiles are farther apart. For example, the figure at right shows the errors which would result from linearly interpolating the ages 15–15.9 female arm-circumference data of Baya Botti et al. [2009], if only the 3%, 10%, 25%, 50%, 75%, 90%, and 97% percentiles were tabulated. Notice that the errors are substantial ($\frac{1}{3}$ cm in arm circumference, almost 1.5% in percentiles) and generally would not average to zero over a sample. This means that, for example, the mean of a sample's linearly-interpolated arm circumference percentiles would generally be systematically lower than the mean of the sample's actual arm-circumference percentiles.



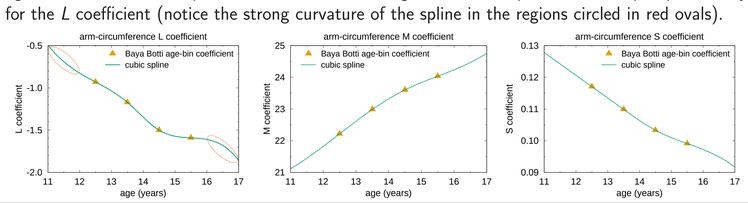
Age Interpolation or Extrapolation

For children and adolescents, anthropometrics' distributions (and hence the distributions' parameters, i.e., mean and SD for a Gaussian or L, M, and S for an LMS distribution) are generally age-dependent. Therefore it's necessary to age-interpolate the parameters for each individual. That is, for each individual whose anthropometrics are to be converted to a z-score:

- Obtain the distribution parameters for each tabulated age bin.
- Interpolate each distribution parameter to the individual's age, assigning each tabulated age bin the bin's average age. (For example, the parameters for a 12–12.9 year age bin should be treated as being for age 12.5.) Cubic spline interpolation generally works well here.
- Use the age-interpolated distribution parameters to calculate that individual's z-score.

If an individual's age is outside the range of tabulated age bins' average ages, the age interpolation becomes an *extrapolation*. Extrapolation is very sensitive to small uncertainties in the data, and to the choice of extrapolant (e.g., the boundary conditions for the cubic spline), so in this case, for each distribution parameter (mean and SD, or L, M, and S), the interpolation/extrapolation function should be plotted along with the parameters for each age bin, so as to verify that the age extrapolation is reasonable.

For example, the figures below show these plots for the Baya Botti et al. [2009] girls' arm circumference data, where the tabulated age bins span the ranges 12.5–15.5 years. The plots show that extrapolation to ages 12 or 16 years is reasonable, i.e., reasonable changes to the extrapolation technique would result in only small changes to the extrapolated L, M, and S coefficients. However, extrapolation to ages 11 or 17 would be quite sensitive to small changes in the extrapolation technique, particularly for the L coefficient (notice the strong curvature of the spline in the regions circled in red ovals).



Endnotes

The `Gaussian_z_of_prob` Function

`Gaussian_z_of_prob`$(P_i/100)$ is defined as the Gaussian z-score corresponding to the percentile P_i. That is, for any probability p, `Gaussian_z_of_prob`(p) is the number z_p such that the area under a Gaussian to the left of $z{=}z_p$ (i.e., the area from $z{=}-\infty$ to $z{=}z_p$) is p. For example, the 25% percentile of a Gaussian corresponds to a z-score of -0.674 (that is, the area under a Gaussian to the left of $z{=}-0.674$, i.e., the area from $z{=}-\infty$ to $z{=}-0.674$, is 0.25), so `Gaussian_z_of_prob`$(0.25){=}-0.674$.

The table at right shows how to compute the `Gaussian_z_of_prob` function in various software environments.

Software	How to compute `Gaussian_z_of_prob`(p)
Microsoft Excel™	`norminv`(p)
Gnuplot	`invnorm`(p)
R	`qnorm`(p)
SciPy (Python)	`from scipy.stats import norm; norm.ppf`(p)

Nonlinear Fitting

Unlike fitting a Gaussian distribution, fitting an LMS distribution requires **nonlinear fitting** (also known as **nonlinear regression**). Nonlinear fitting is now widely available in software packages such as Microsft Excel™, Gnuplot, R, and SciPy (Python). Nonlinear fitting is an iterative process: the user supplies an "initial guess" which the curve-fitting software then repeatedly adjusts to fit the data points better and better. Our LMS-fitting technique uses a Gaussian approximation to compute the initial guess.

Mathematical Note: Note that the "nonlinear" in "nonlinear fitting" does not refer to whether the fitted curve is straight or curved. Rather, "nonlinear" refers to how the y coordinate of a point depends on the curve's parameters (the mean and SD for a Gaussian, or L, M, and S for an LMS distribution). For an LMS distribution this dependence is given by equation (1), which depends nonlinearly on L and S.

References

A. Baya Botti, F. J. A. Pérez-Cueto, P. A. Vasquez Monllor, and P. W. Kolsteren. Anthropometry of height, weight, arm, wrist, abdominal circumference and body mass index, for Bolivian adolescents 12 to 18 years – Bolivian adolescent percentile values from the MESA study. *Nutrición Hospitalaria*, 24(3):304–301, 2009. ISSN 0212-1611.

T. J. Cole. Fitting smoothed centile curves to reference data. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 151(3): 385–406, 12 1988.

T. J. Cole. The LMS method for constructing normalized growth standards. *European Journal of Clinical Nutrition*, 44(1):45–60, 1990.

T. J. Cole. The development of growth references and growth charts. *Annals of Human Biology*, 39(5):382–394, 2012.

T. J. Cole and P. J. Green. Smoothing reference centile curves: The LMS method and penalized likelihood. *Statistics in Medicine*, 11:1305–1319, 1992.