Using a Half-Normal Distribution to Track Extreme BMI Values Among Youth

Rong Wei, Van Parsons, and Cynthia Ogden

National Center for Health Statistics, CDC

Key word: body mass index; extreme BMI, z-score; growth percentiles; half-normal distribution.

1 Introduction

1.1 Background

The 2000 CDC growth charts (Kuczmarski, et al., 2002) (henceforth referred to as Growth Charts) provide charts and models that are used extensively in clinical practice to assess growth or patterns of child growth, and in the US are also used to define obesity (Ogden CL, Flegal KM. 2010). The models and charts presented in the Growth Charts are based on national data collected between 1963 and 1994 and provide references useful for tracking US child growth. The Growth Charts include sex-and age specific body mass index (BMI) percentiles, for children and adolescents aged 2-19 years, between the 3rd and 97th values along with three normal transformation parameters (lambda, mu and sigma, LMS), which in turn can be used to interpolate other percentiles and associated z-scores (Flegal and Cole, 2013; Kuczmarski et al., 2002). Extrapolating beyond the 97th percentile by using the provided Growth Charts LMS parameters is not recommended because there was insufficient data beyond the 97th percentile to model additional percentiles. Z-scores obtained using extrapolated LMS parameters are compressed so that large changes in extreme BMI values reflect small changes in z-scores (Woo, 2009).

1.2 Objectives

For simplicity of terminology, the individual BMI growth charts contained in the Growth Charts will be referred to as the "GCharts". The important measure "*obesity"* is defined as a BMI at or above the sex-and age specific 95th percentile (*BMIp95*). However, since the publication of the Growth Charts, the prevalence of obesity among youth has increased from 5% in 1976-1980 to 18.5% in 2015-2016 (Hales et al., 2017). Given this order-of-magnitude of change, and the GChart limitations for tracking extreme BMI values, a need has arisen to create a complementary sex- and age specific metric suitable for tracking extreme BMI. This paper presents details of an exploratory study of a proposed methodology based on the half-normal distribution to model extreme BMI values.

2 Statistical model

2.1 Population and Data

The GCharts are based on modeled aggregate populations built on NHANES data collected

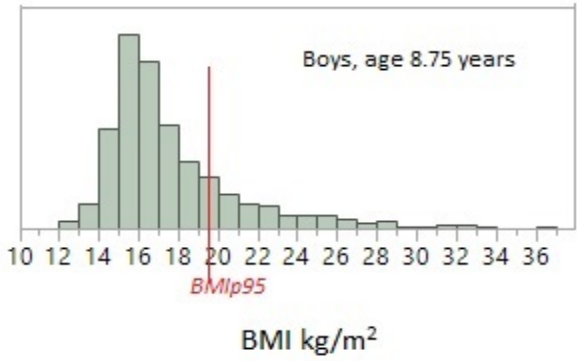
Disclaimer: The findings and conclusions in this study are those of the authors and do not necessarily represent the views of the Centers for Disease Control and Prevention.

from 1963-1994. In that time frame, the tail distributions had sparse data. Now, with 18 additional years of NHANES data available (1999-2016), a total of 8777 observations for children age 2 to 19 years greater than a GChart value BMIp95 are available for modeling.

These data are grouped by each sex into 36 six-month-age groups for model fitting. While the original GCharts are based on somewhat involved models, for fitting a complementary tail-distribution model, the authors decided to fit and assess a one-parameter half-normal distribution model applied to each of the 72 groups. This model was selected as it provides straight-forward formulas for BMI percentiles or z-scores, and would be easy to incorporate into a web page and adapted BMI charts targeted for users who need to track extreme BMI values.

Figure 1 shows the BMI distribution for boys in age group around 8.75 years for the original growth charts data along with the additional data from 1999-2016. The full range of BMI values is skewed toward the right. The red straight line represents the obesity threshold or GChart BMIp95. It will be those points to the right of the red line that will be fitted by a one-parameter half-normal model.

Figure 1. BMI distribution 1963 to 2016 with BMIp95 threshold.



2.2 The model

The current task is to only focus on BMI values greater than sex- and age-specific *BMIp95* and to create a reasonable data-driven model for the tail portion.

This can be done by first recalling the following conditional probability structure of a random variable B:

$$P(B \geq t_{\text{pth}}) = P(B \geq t_{\text{pth}} \mid B \geq t_{95\text{th}}) \cdot P(B \geq t_{95\text{th}}), \quad (1)$$

where t_{pth} is the p^{th} percentile that satisfies the order $t_{pth} > t_{95th}$, and B = BMI for this study.

The conditional probability term, $P(B \geq t_{\text{pth}} \mid B \geq t_{95\text{th}})$, represents the tail distribution; its form is unknown and will be based on a conjectured and then validated model. The last multiplicative term in (1) properly links the tail distribution with the pre-tail part of the original distribution.

2.3 The half-normal distribution and properties

For age group a it will be assumed that the tail distribution can be modeled by a half-normal distribution. This distribution has density:

$$2\frac{1}{\sqrt{2\pi}\sigma_a}e^{-x^2/(2\sigma_a^2)}, x > 0 \quad (2)$$

where σ_a is a distribution scale-parameter that is proportional to both this distribution's mean and standard deviation.

As a theoretical model, a scaled expectation for distribution (2) is

$$\sqrt{\frac{\pi}{2}}E(X) = \sigma_a \quad (3),$$

and the second moment is $E(X^2) = \sigma_a^2$

The half-normal probability distribution can be expressed in terms of the standard normal distribution, $\Phi(x)$. The half-normal distribution satisfies:

$$P(X > x) = 2(1-\Phi(\frac{x}{\sigma_a})), \text{ for } 0 < x < \infty. \quad (4)$$

Returning discussion to the study with B = BMI, the tail region $\{B > t_{95th}\}$ is assumed to have the
conditional probability distribution, $P(B \leq t_{pth} \mid B > t_{95\text{th}})$, specified by a half- normal distribution.
Re-expressing equation (1) in cumulative distribution function (CDF) form, along with expression (4) in
CDF form, the CDF of B can be expressed:

$$P(B \leq t_{pth}) = 1-[1-2(1-\Phi(\frac{x}{\sigma_a})]\cdot(0.05) \quad (5),$$

whenever t_{pth} is the p^{th} percentile of B, $t_{pth} > t_{95th}$, and x is defined as $t_{pth} = t_{95th} + x$, $x > 0$.

For $t_{pth} < t_{95th}$ the CDF values are those specified by the known pre-tail CDF.

As an example of usage, given a p (or percent) value $p \geq 0.95$, the corresponding percentile value, $t_{pth} = t_{95th} + x$, can be solved as

$$t_{pth} = t_{95th} + x^*, \text{ where } x^* = \sigma_a \cdot \Phi^{-1}\left(\frac{1}{2}\left(1+\frac{p-0.95}{1-0.95}\right)\right), \quad (6)$$

where Φ^{-1} is the inverse function of Φ.

If a "z-score-scale" unit rather than a "p-level scale" unit in the tail region is provided, i.e., $z > 1.64$ (assuming cases where the percentile associated with z is greater than the 95th), the corresponding p in formula (6) will be $p_z = \Phi(z)$. This p_z value can be substituted (6) to obtain the BMI value corresponding to z, i.e.,

$t_{pth} = t_{95th} + x^*$, where $x^* = \sigma_a \cdot \Phi^{-1}\left(\frac{1}{2}\left(1 + \frac{\Phi(z) - 0.95}{1 - 0.95}\right)\right)$, (7)

2.4 Applications of half-normal tail distributions for GCharts

Once an estimate of σ_a is established, the rules of (5), (6) and (7) can be used to define tail metrics for the GCharts.

a. Given a BMI value consistent with obesity and defining x = (BMI - *BMIp95*), equation (5) determines the percentile.
   A new measurement determined to be in a sex-age specific obese range can have its percentile quantified relative to its GChart standard.
b. Given a value of p or z, equations (6) and (7) determine the corresponding BMI value, i.e., the value of t_{pth}.

   If the BMI value that exceeds the p^{th} percentile (e.g., 99%) is needed that value can be determined. A set of tracking curves defined by obesity percentiles can be created.

3. Estimation of σ_a

3.1 Finite population sampling approach

The selected half-normal model is fully specified by one-parameter, σ_a. As the NHANES data are based on a complex-survey design, finite-population-based sampling methods will be used. An assumption will be made that the sampling population's tail shape for group a is closely approximated by a shape consistent with equation (2). (This assumption is in the spirit of imposing population regression parameters on a finite population). The estimators of the finite population forms of (3), $\sqrt{\frac{\pi}{2}}E(X)$ or $\sqrt{(E(X^2))}$, are candidates for defining an estimator of σ_a.

The usual finite population estimator of a population mean is the sample-weighted mean. Sample-weighted means on observed data y or y^2 on the specified tail of the population are the two candidate forms. An empirical study led to choosing the y-form. More precisely, for the NHANES data for each group a and individual i, we let w_i be the survey weight and $y_i = (bmi_i - BMIp95)$ be the BMI component exceeding its GChart's 95th percentile; all those units that do not exceed are dropped.

Letting n_a be the sample size in the tail, the finite population estimator of the parameter σ_a of expression (3) is defined as

$$\hat{\sigma}_a = \sqrt{\frac{\pi}{2}} \frac{\sum_{i=1}^{n_a} w_i y_i}{\sum_{i=1}^{n_a} w_i} \quad , (8)$$

3.2 Small samples and reducing impact of outliers

The estimator in expression (8) is subject to outlier influence, especially for smaller sample sizes. As the data are based on several cycles of NHANES, the nominal sample sizes n_a are greatly reduced by variation of sampling weights and clustering within each group. Incorporating the design structures of 50 years of data collection is beyond the scope of the current research. A commonly used technique, (Kish 1992, Henry and Valliant 2015), is to use survey weight variation to reduce the nominal sample size to an *effective sample size*. More precisely, if an age group a has a nominal sample size n_a, its effective sample size will be a reduction by a factor $deff = (1 + cv^2(\text{weights}))$, and $n_{a.eff} = \frac{n_a}{deff}$. For this study, most computed $deff$'s were in the range of 2.0 in magnitude. This value was used for all groups and for all calculations discussed in this document.

3.3 Extreme outliers

Robust variations of the estimator $\hat{\sigma}_a$ can be defined to give less weight to any extreme y values in equation (8). Our attempt is to find <u>one</u> functional form that works well with all 72 sample sex-age groups. An iterative trimming of the extreme y 's is implemented to keep the $\text{trimmed}(y) \le k\hat{\sigma}_a$ for a selected value k.

Algorithm:

Fix a k

Compute $\hat{\sigma}_a$ as in (8), then compute $z = \frac{y}{\hat{\sigma}_a}$, next define

$y_{\text{new}} = y_{\text{old}}$ if $z < k$, and $y_{\text{new}} = k$ otherwise.

Iteratively, recompute $\hat{\sigma}_a$ using (8) with the value y_{new} until convergence.

Call the new estimate $\hat{\sigma}_{a.k}$

Three $\hat{\sigma}_a$ *forms* were considered to assess the impact of restricting the outliers to a mutiplicative factor of the computed $\hat{\sigma}_a$: original $\hat{\sigma}_a$, trimmed at $k = 2$: $\hat{\sigma}_{a.2}$, and trimmed at $k = 3$: $\hat{\sigma}_{a.3}$. The y 's were only trimmed in the computation of $\hat{\sigma}_a$, not for other computed statistics

4 Smoothing model $\hat{\sigma}_a$'s across sex-age groups

The procedures of section 3 provide 72 values of $\hat{\sigma}_a$ for each original or trimmed variation. A goal of this project was to provide simple sex-specific computational expressions available for users of GCharts who want to interpolate simple data input for a given ages on the continuum of 2 to 19 years. Smoothing $\hat{\sigma}_a$ across age groups with a parametric regression to obtain a simple expression of $\hat{\sigma}_a$ as a function of age will provide a means to do that.

4.1 Polynomial regressions

Linear and quadratic smoothers of the $\hat{\sigma}_a$'s of the forms

Linear: $\hat{\sigma}_a = c_0 + c_1 a + e$

Quadratic: $\hat{\sigma}_a = c_0 + c_1 a + c_2 a^2 + e$ were sought.

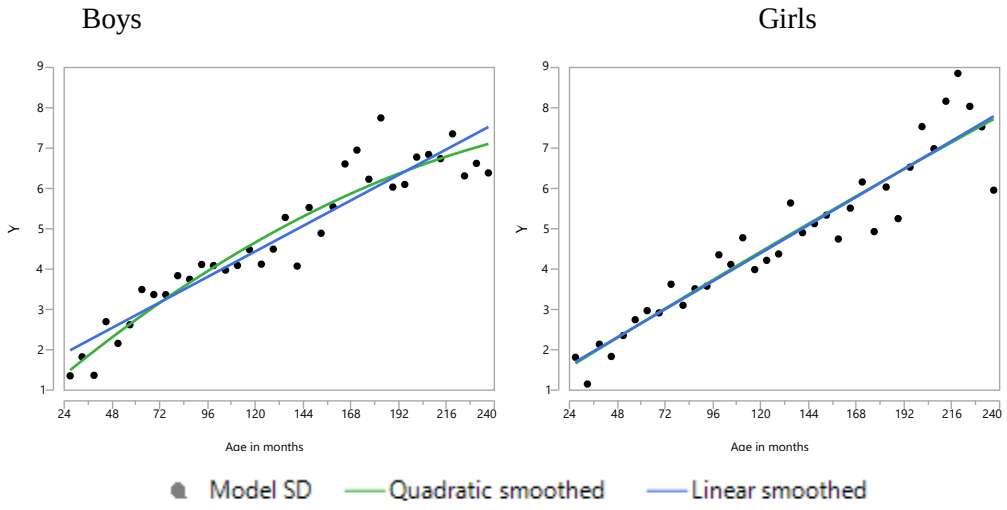
The single sex and age group estimates of σ_a were smoothed using fitting both linear and quadratic polynomials over the 36 age groups on for each sex. Weighted least squares (WLS) procedures were used.

Three options for WLS weights were considered:

1. $LSwt_a = n_{a.eff}$ as discussed in 3.2
2. $LSwt_a = \dfrac{n_{a.eff}}{\hat{\sigma}_a^2}$
3. $LSwt_a = \dfrac{n_{a.eff}}{\text{iterative fitted } \hat{\sigma}_a^2}$

Here, the assumption is made that large sample sizes and/or small $\hat{\sigma}_a^2$ is related to increased precision. The $LSwt$ option of 3. above replaces the value $\hat{\sigma}_a^2$ with the fitted polynomial at point a and then repeats the WLS until convergence. In the situation at hand, for a selected $\hat{\sigma}_a$, $\hat{\sigma}_{a.2}$ or $\hat{\sigma}_{a.3}$ and for selected linear or quadratic fitting there was no practical difference in the fits using any of the three methods. The choice between a linear and quadratic fitting can be made by examining the R^2 statistic and examining scatterplots. **Figure 2** shows a scatterplot of the observed $\hat{\sigma}_a$ and fitted linear and quadratic values $\hat{\sigma}_{a.smoothed}$ for both the boys and girls. The quadratic form fitted both boys and girls reasonably well and was selected. The notation $\hat{\sigma}_{a.smoothed}$ will be used to denote this estimator.

Figure 2. $\hat{\sigma}_{a.smoothed}$ linear and quadratic curves vs. original model estimated $\hat{\sigma}_a$



5 Assessments of the half-normal fits

5.1 CDF plots as a diagnostic

Several estimation methods for fitting a one-parameter half-normal model have been explored. As the data are based on complex survey samples covering a 50-year span, traditional model assessment based on independent sampling models do not strictly apply. For this study, assessments are made by visual inspection and aided by some commonly practiced techniques that treat survey units like independent sample units.

One straight-forward model assessment is to compare the CDF of the half-normal model fitted with an estimate $\hat{\sigma}_a$ to the empirical survey-weighted CDF.

For each of the 72 sex-age groups, a weighted empirical distribution function, $\hat{F}_a(x)$, was computed using the R Package "Hmisc" and function "wtd.Ecdf" on BMI values that exceeded BMIp95. For each $\hat{F}_a(x)$ an empirical 95% confidence interval (point-wise) was computed by using a standard error estimate:

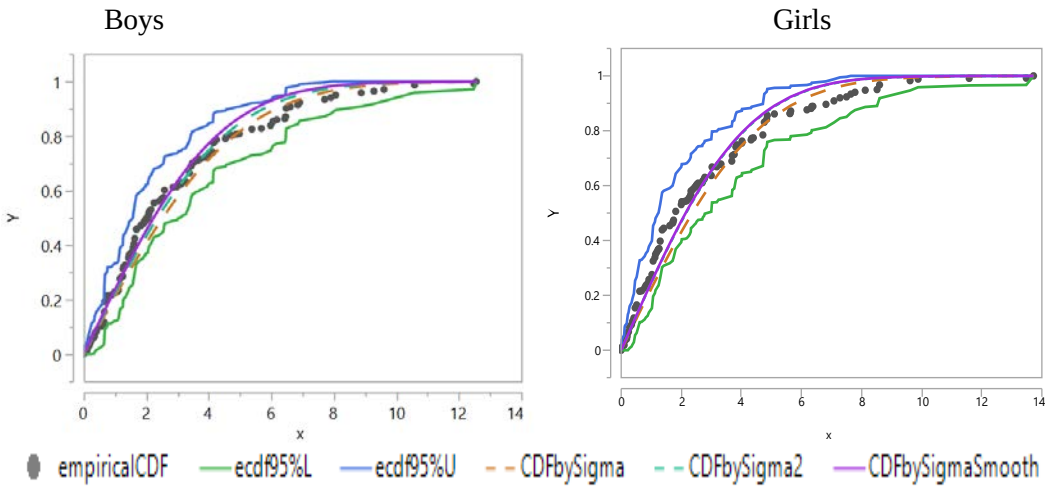
$$sd_{F_a} = \sqrt{\frac{\hat{F}_a(x)(1-\hat{F}_a(x))}{n_{a.eff}}}$$, $n_{a.eff}$ = effective sample size for group a (see sect 3.2), and

then treating $\hat{F}_a(x) \pm 2sd_{F_a}$ as an approximate 95% confidence interval.

If a sex-age group distribution is indeed a half-normal with parameter σ_a , then equation (4) can also be used to model the CDF.

A goodness of fit (GOF) is established visually by examining the degree to which the model (4) CDF is contained in the empirical cdf. **Figure 3** shows overlayed half-normal fits for $\hat{\sigma}_a$ equal to the raw fitted parameter, the trimmed parameter at $k = 2$ $\hat{\sigma}_{a.2}$ (see section 3.3), and the quadratic smoothed $\hat{\sigma}_{a.smooth}$ parameter.

Figure 3. Examples of diagnostic from 72 sex-age specific CDF plots after half-normal modeling - empirical CDF with 95%CI vs. CDFs by $\hat{\sigma}_a$, $\hat{\sigma}_{a.2}$ and $\hat{\sigma}_{a.smooth}$



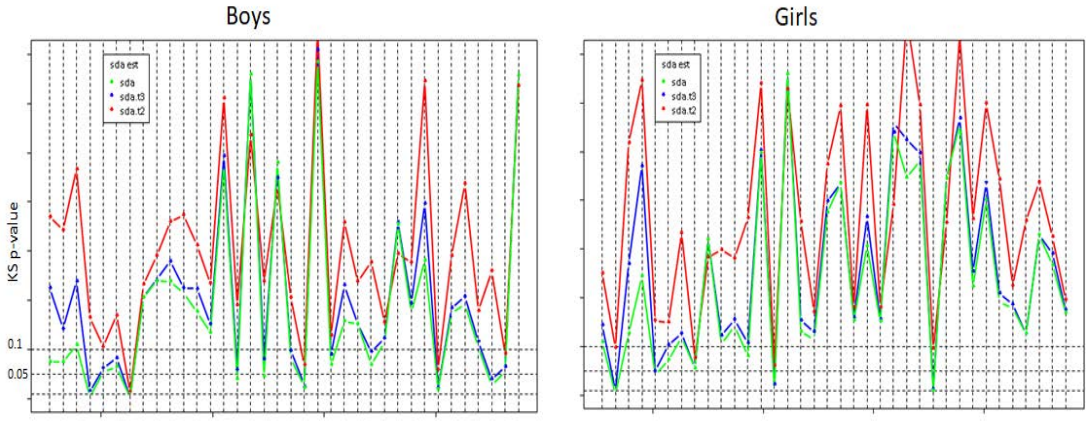
5.2 Kolmogorov-Smirnov goodness of fit test

The Kolmogorov-Smirnov goodness-of-fit (KS-GOF) test is intended for testing an independent identically distributed - based sample against a completely specified CDF. The half-normal with an estimated $\hat{\sigma}_a$ does not satisfy the NULL distributional assumptions of the KS-GOF, but the computation of KS-GOF with an estimated parameter does give an indication of fit, especially when comparing among the different methods for estimating σ_a.

For each sex-age group and estimation method, the R package "ks.test" was used to produce a *p*-value, and to adjust for the multiple comparisons of 72 age-groups, the R package "p.adjust" was used. The R package had options for 5 different multiple comparison methodologies, all of which produced a *p*-value ≥ 0.05 for the original non-modified estimate of $\hat{\sigma}_a$.

Figure 4 provides a plot of the stand-alone p-values for the fitted half-normal based on parameters $\hat{\sigma}_a$, $\hat{\sigma}_{a.3}$ and $\hat{\sigma}_{a.2}$. As can be seen, the *p*-values for the truncated σ's tend to be larger than the original raw version. For $\hat{\sigma}_a$, $\hat{\sigma}_{a.3}$ and $\hat{\sigma}_{a.2}$,10, 6, and 1 groups, respectively, out of 72 total groups were significant at $p < 0.05$ stand-alone level. These results indicate that better fits may occur whenever some trimming of the larger observed BMI values is implemented when estimating $\hat{\sigma}_a$

Figure 4. Stand-alone p-values of 72 Kolmogorov-Smirnov goodness of fit tests



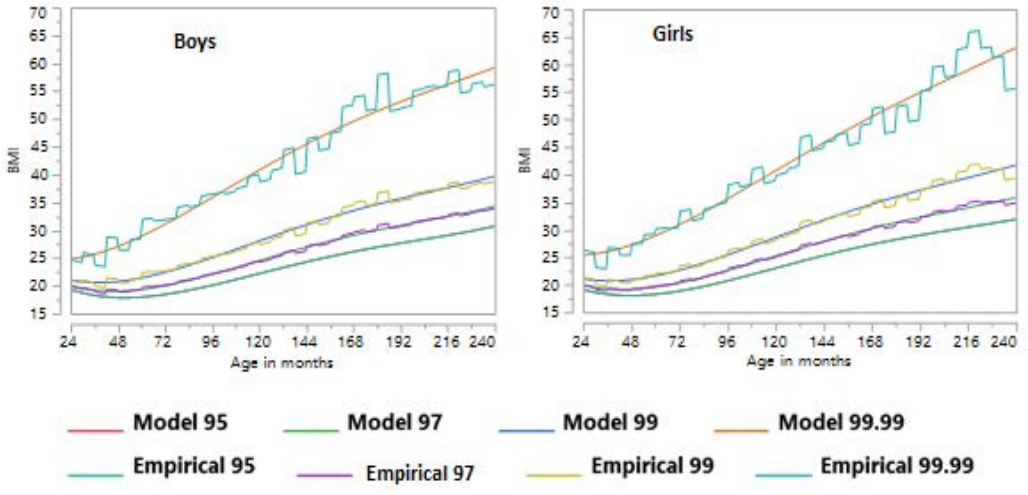
Kolmorgorv-Smirnov test for each fit*
*Indication of fit as the fitted model is based on an estimated parameter

5.3 Comparisons with BMI tail data.

In the examples that follow, the proposed $\hat{\sigma}_{a.smooth}$ is considered as generating the half-normal distribution and examined with respect to the 1963-2016 NHANES BMI tail data (values that exceed BMIp95).

In **Figure 5** the empirical CDF tail percentiles at values 95, 97, 99, and 99.99 are plotted at 36 age points and overlaid with the $\hat{\sigma}_{a.smooth}$ generated half-normal percentiles at the corresponding ages. The 95-99 percentiles appear to fit the empirical data well (at least visually). The 95 value is a perfect fit by definition (see formula (5)). The 99.99 value is include to examine an extreme case of possible fit deviation. Here, in this case, the sparsity of data in the fitting process may lead to less than optimal fitting.

Figure 5. Plot comparisons with data, empirical, model/smoothed percentiles
Projected percentile curves using smoothed curves based on $\hat{\sigma}_{a.smooth}$ compared to empirical percentiles by sex-age groups



6 Discussion

The Growth Charts are a resource for providing references for tracking child growth. However, the Growth Charts are based upon 1963-1994 data, and there was sparse data available beyond the 97th percentile. As child obesity in the US has increased, extreme percentiles beyond the 97th percentile are increasingly relevant for tracking extreme BMI values. In the years since the publication of the Growth Charts, the NHANES has collected additional data on BMI among children and adolescents 2-19 years. These data, along with the original growth charts data, can be used to create complementary metrics for extreme BM values that to compensate for the sparse data beyond the 97th percentiles in the Growth Charts. This study suggests using a one-parameter half-normal model for fitting BMI tail probabilities/percentiles.

Summary of research:

1. For each sex-age group a, a single parameter, σ_a, can be used to completely specify the BMI distribution for obese children. This parameter, in its basic form, can be estimated using a survey-weighted mean. The suggested estimate can be trimmed in small sample situations to reduce influence of outliers.
2. The basic estimate of σ_a is smoothed and expressed as a simple quadratic function of age to be amenable to lay users of the GCharts.
3. NHANES design features of survey weights and effective sample sizes are used in the estimation procedures.
4. The fit of the model is assessed by
    a. Examination of empirical and modeled CDFs and percentiles.
    b. Multiple-comparisons of heuristic Kolmogorov-Smirnov goodness-of-fit statistics.

This proposed half-normal method appears to fit the current data reasonably well and can be used to track extreme BMI values. Additional material is contained in Wei et al. (2019).

7 References

Flegal KM, Cole TJ. Construction of LMS parameters for the Centers for Disease Control and Prevention 2000 growth charts. Natl Health Stat Report. 2013 Feb 11;(63):1-3.

Hales CM, Carroll MD, Fryar CD, Ogden CL. Prevalence of Obesity Among Adults and Youth: United States, 2015-2016. NCHS Data Brief. 2017 Oct;(288):1-8.

Henry, KA, Valliant, RL. 2015. A Design Effect Measure for Calibration Weighting in Single-Stage Samples. Survey Methodology, 41(2), 315-331.

Kish L 1992 . Weighting for unequal Pi. *Journal of Official Statistics*, 8, 183-200

Kuczmarski RJ, Ogden CL, Guo SS, Grummer-Strawn LM, Flegal KM, Mei Z, Wei R, Curtin LR, Roche AF, Johnson CL. 2000 CDC Growth Charts for the United States: methods and development. Vital Health Stat11 2002(246):1-190.

Ogden CL, Flegal KM. 2010. Changes in terminology for childhood overweigh and obesity. National health statistics reports; no 25. Hyattsville, MD.

Wei R, Ogden CL, Parsons V, Freedman D, Hales C. 2019. Redefining BMI z-score and percentile values above the 95th percentile of the CDC growth charts. In progress.

Woo JG. Using body mass index Z-score among severely obese adolescents: a cautionary note. Int J Pediatr Obes. 2009;4(4):405-10.