Journal of Applied Statistics

ISSN: 0266-4763 (Print) 1360-0532 (Online) Journal homepage: http://www.tandfonline.com/loi/cjas20

Redefining standards for body mass index of the US population based on BRFSS data using mixtures

Tatjana Miljkovic, Saleem Shaik & Dragan Miljkovic

To cite this article: Tatjana Miljkovic, Saleem Shaik & Dragan Miljkovic (2016): Redefining standards for body mass index of the US population based on BRFSS data using mixtures, Journal of Applied Statistics, DOI: 10.1080/02664763.2016.1168366

To link to this article: http://dx.doi.org/10.1080/02664763.2016.1168366

Published online: 15 Apr 2016.

JOURNAL OF APPLIED STATISTICS, 2016
http://dx.doi.org/10.1080/02664763.2016.1168366

Taylor & Francis
Taylor & Francis Group

Redefining standards for body mass index of the US population based on BRFSS data using mixtures

Tatjana Miljkovic[a], Saleem Shaik[b] and Dragan Miljkovic[b]

[a]Department of Statistics, Miami University, Oxford, OH, USA; [b]Department of Agribusiness & Applied Economics, North Dakota State University, Fargo, ND, USA

ABSTRACT

Using body mass index (BMI) data from 2012 Behavioral Risk Factor Surveillance System, we test a spectrum of single parametric skewed distributions as well as Gaussian mixture densities to determine best distributional fit. We find that a k-component Gaussian mixture is the best model to describe the distribution of BMI data for the overall US population and for the population divided by gender, race, and region. A 4-component Gaussian mixture with the following subpopulation means (standard deviations) fits best the US population: 22.21($\sigma = 2.27$), 26.05($\sigma = 2.19$), 29.83($\sigma = 3.90$), 35.47($\sigma = 8.45$) with corresponding weights: 23%, 25%, 37%, and 15%. Current obesity standards are set based on a convention and they are fairly dated. Overweight population has BMI (25.0, 29.9). Obese population is subdivided into three grades based on BMI: grade 1 (30–35), grade 2 (35–40), grade 3 (40 and above). Our study shows that modeling BMI using mixtures can be used to redefine current standards and support them with actual prevalence rather than a dated convention. By redefining BMI standards and employing the mixture models by gender and race, health and food policy makers will have opportunity to diversify policies and treatments of obesity as premier public health problem in the USA.

ARTICLE HISTORY
Received 12 December 2014
Accepted 16 March 2016

KEYWORDS
Obesity; BMI; mixtures

AMS SUBJECT CLASSIFICATION
6204; 6207; 62N02; 62P10; 62P20

1. Introduction

Obesity has been one of the most important public health issues in the USA in recent decades. A common measure of obesity is a body mass index (BMI) of 30 or greater. BMI is calculated as weight in kilograms divided by height in meters squared. Based on recommendations of the panel of experts dated back to 1998 [11], which are still used as the standard, overweight is defined as a BMI ranging from 25.0 to 29.9. The same standard defines obesity as a BMI of at least 30.0 and subdivided into three grades: grade 1 with $30.0 \leq \text{BMI} < 35.0$; grade 2 with $35 \leq \text{BMI} < 40.0$, and grade 3 with $\text{BMI} \geq 40.0$.

Most studies [12–16,19,31], that address the prevalence of obesity in the USA with its important public health and health policy implications use the BMI data from the

CONTACT Tatjana Miljkovic miljkot@miamioh.edu Department of Statistics, Miami University, Oxford, OH 45056, USA

National Health and Nutrition Examination Surveys (NHANES), or, alternatively, from the Behavioral Risk Factor Surveillance System (BRFSS) [9,24, 25].

The first of NHANES nationally representative health examination surveys of civilian non-institutionalized population using a complex, stratified, multi-stage probability cluster sampling design was conducted in 1960. In 1999, NHANES became a continuous survey with data released in 2-year cycles [15]. BRFSS is a CDC-sponsored, state-based telephone survey of health risk factors with the main purpose to provide state-specific estimates of the prevalence of behaviors that are associated with the leading causes of death in the USA. Each participating state independently selects for interview a probability sample from adult residents who are at least 18 years old in households with telephones. All states, during the same year, use an identical core questionnaire administered over the phone by trained interviewers. Questions about height and weight, based on which BMI is calculated, are asked, among others, by the interviewers. When the two data sets are compared, it has been determined that the prevalence estimates of overweight and obesity generated by the BRFSS under-estimate those from the NHANES [36]. While both data sets have been widely used, they both have some limitations. For example, NHANES does not sample an adequate number of persons who are members of racial/ethnic minority communities other than non-Hispanic blacks and Mexican-Americans to permit estimating obesity prevalence in these communities (refer to [6]).

The National Health Interview Survey (NHIS), which is conducted by the National Center for Health Statistics of the Centers for Disease Control and Prevention (CDC), has been the main source of national health data on the US population since the 1950s [34]. The NHIS obtains information on a variety of health measures, including medical conditions, access to health care, and health risk factors, and plays a pivotal role in tracking national health objectives. Data are collected in a centralized manner by the US Bureau of the Census via within-household, in person interviews. Information is obtained from a nationally representative sample of adults and children, and both self-reports and proxy data are included.

The issue of comparability between these data sources is of more than academic interest [2]. Telephone-administered surveys such as the BRFSS have substantial cost and timeliness advantages over household-administered surveys such as the NHIS which have the advantage of being able to collect more detailed information on a wider range of topics [17,18]. As already mentioned, previous studies have shown that self-reports underestimate weight, McLachlan and Peel [23] and data from the NHANES, which obtains measured height and weight from respondents, demonstrate that self-reports substantially underestimate BMI [21]. Thus, both the NHIS and the BRFSS are likely to substantially underestimate the extent of overweight and obesity, with the underestimate for obesity being larger in the BRFSS. However, BRFSS data can be combined across states to provide national estimates for certain measures and to produce estimates generally comparable to those of the NHIS, although there may be differences for subgroups. The importance of the differences in estimates between the two surveys will depend on the purposes and situations for which these estimates are to be used. In addition, the BRFSS could provide rapidly available data for helping guide national policy and program decisions in areas such as tobacco control, insurance coverage, and obesity. For example, although BRFSS obesity estimates are lower than those of the NHIS or NHANES, the BRFSS has provided the most timely state and national data demonstrating the worsening of US obesity trends over the

past decade [26]. Finally, use of BRFSS data in computing national estimates for selected measures has the added benefit of improving the comparability of state and national data, in that the two types of estimates would be based on the same questions and mode of interview [27].

Important previous studies on prevalence of obesity and trends in the USA that used NHANES BMI data levels in the USA use logistic regression of the BMI data on one (or more) explanatory variables with linear trend over time [12–16]. Likewise, many studies that used BRFSS BMI data to examine causes of obesity and its health, economic, and policy implications, also used logistic regression, and logit or probit statistical models [4,10,24,25,32,35]. Most of these studies make the assumption that BMI is binary (BMI $\geq 30 = 1$ is obese and BMI $< 30 = 0$ is not obese), and assume binomial probability distribution, for given values of the explanatory variables. This is a greatly simplifying assumption that hides the important information on the true distribution of the BMI. Moreover, it is based on a convention. For instance, BMI of 25 kg/m^2 is used as the proposed upper limit of normal, although others could be selected based on ethnicity, age, or other considerations [5]. Hence it is clear that these defining limits are, albeit based on scientific and policy experts' convention, arbitrary. While it is clear that logistic regression provides a useful means for modeling dependence and determining correlations of a binary response variable with one or more explanatory variables [3], it neither describes the prevalence of obesity nor determines causality of it. Thus determining true probability distribution of BMI is important not only in order to accurately describe it in statistical terms but also to be able to interpret as being derived from an underlying set of other random variables.

The aim of this study is to investigate the true distribution of the BRFSS BMI survey data for (1) the overall American population (2) regional BMI distribution where the regions are defined based on the CDC obesity prevalence rates; (3) subpopulation based on gender and race. Within each population, mixture analysis is used to investigate whether a mixture of two (or more) normal (or other) distributions explain the variance in BMI better than a single distribution.

In probability and statistics, a mixture distribution is the probability distribution of a random variable (BMI in this case) whose values can be interpreted as being derived in the following way from an underlying set of other random variables: specifically, the realization of the random variable with a mixture distribution is randomly selected from among the realizations of the underlying random variables, with a certain probability of selection being associated with each. Here the underlying random variables may be random vectors (each having the same dimension) in which case the mixture distribution is a multivariate distribution.

Mixture distributions arise in many contexts in the literature and arise naturally where a statistical population contains two or more subpopulations, as is the case of BMI. They are sometimes used as a means of representing non-normal distributions, which we will test if that is the case too. The advantage of testing for and using mixture distributions is, primarily, because parametric statistics that assume no error often fail on such mixture densities. For example, statistics that assume normality often fail disastrously in the presence of even a few outliers. The analysis should be conducted with robust statistics.

The organization of this paper is as follows. Section 2 defines the methodology used. Section 3 describes the data and the analysis. Conclusion is provided in Section 4.

2. Methodology

For many years, prior to year 2000, the distribution of BMI was assumed to follow a normal distribution or approximately a bell shaped curve as this was similar case for measurements of weight and height. Since this assumption has not been clearly proven or documented, first we tested the normality of the BMI data using the Jarque-Bera normality test and D'Agostino's K-squared tests.

Recently, some claims were made that BMI is not normally distributed [29] indicating that a Log-normal or some other positively skewed distribution may be a better fit. Hence, the following single distributions, exhibiting a positive degree of skewing in the right tail of the data, were fitted using the method of maximum likelihood:

- Log-normal(μ, σ) where mean is μ and standard deviation is σ.
- Gamma(θ, α) where shape is $\theta$and scale is α.
- Logistic(μ, s) where mean is μ and scale is s.
- Inverse Gaussian(μ, θ) where mean is μ and shape is θ.
- Weibull(θ, τ) where shape is θ and scale is τ.

Optimization of the non-differentiable likelihood functions, for single component modes, was built on Nelder–Mead algorithm.

Finally, a univariate Gaussian mixture model was tested to allow for inclusion of different sub-populations while fitting the overall population of BMI data. Finite Gaussian mixture modeling was founded on the EM algorithm, well known for its applications in model-based clustering and classifications [23]. For the observations $y_1, y_2, \ldots, y_n$, the mixture density function is given by

$$f(y_i|\theta) = \sum_{k=1}^{K} \pi_k \phi_k(y_i|\mu_k, \sigma_k^2),$$

where the kth mixing proportion π_k represents the probability that observation y_i belongs to the kth subpopulation with corresponding kth normal component density $\phi(.)$ Here, $\theta = (\mu, \sigma^2, \pi)$, $\mu = (\mu_1, \mu_2, \ldots, \mu_K)$, $\pi = (\pi_1, \pi_2, \ldots, \pi_{K-1})$ with $\sum_{k=1}^{K} \pi_k = 1$, and $\phi(y|\mu, \sigma^2)$ is the normal distribution with mean μ and variance σ^2 evaluated at y.

Therefore,

$$f(y|\theta) = \prod_{i=1}^{n} \sum_{k=1}^{K} \pi_k \phi_k(y_i|\mu_k, \sigma_k^2).$$

Define latent indicators $z_1, z_2, \ldots, z_n$, such that $z_i \in (1, \ldots, K)$ and $p(z_i = k|\theta) = \pi_k$, the augmented model for (y, z) is defined by the following joint density

$$f(y, z|\theta) = \left[\prod_{k=1}^{K} \prod_{i \in I_k} \phi_k(y_i|\mu_k, \sigma_k^2) \right] \prod_{i=1}^{n} p(z_i|\theta),$$

where $I_k = (i : z_i = k)$.

The best model was selected based on Akaike information criterion (AIC) [1], Bayesian information criterion (BIC) [30], and Kolmogorov–Smirnov (K–S) [22] test statistic

results. For each component in the mixture the estimates of mixing proportions, mean, and standard deviations were summarized. Statistical analysis was conducted using R version 3.1.0 (The R Foundation of Statistical Computing [33]). The following libraries from R were utilized: MASS, stats4, mclust, tseries, mixtools, splines, and statmod).

3. Application

3.1. Data

Data is obtained from the BRFSS. The BRFSS is conducted by the National Center for Health Statistics of the CDC [7]. According to CDC,

> BRFSS is the nation's premier system of health-related telephone surveys that collect state data about US residents regarding their health-related risk behaviors, chronic health conditions, and use of preventive services. BRFSS [8] collects data in all 50 states as well as the District of Columbia and three US territories. BRFSS completes more than 400,000 adult interviews each year, making it the largest continuously conducted health survey system in the world. (CDC [7], http://www.cdc.gov/brfss/)

The BMI information along with state, county, age, gender, and racial/ethnic-specific category variables from the 2012 BFRSS survey are used in this analysis. Prior to conducting the analysis, the presence of aggregate district, state and regional data, that is, observations with county code 777, 888 and 999 are deleted to avoid bias in the distributions. In addition, observations with income source code 77 (do not know) and 99 (refused) were deleted. Data for Alaska Native were not available. Finally, observations from the top five race groups - White, Black or African American, Asian, Native Hawaiian or Other Pacific Islander and American Indians are used in the analysis. This included interracial observations within the top five race groups. Only adults age 18 and above were included.

The US county map of the average BMI by county is presented in Figure 1. This map shows visually the spatial distribution of average BMI values. While it is a common practice by CDC to show a state map by region based on prevalence rate (http://www.cdc.gov/obesity/data/adult.html), we believe that the state map does not capture serious gap across regions of the state and among racial/ethnic groups; therefore, county map may be used as a better visual tool. In Figure 1, for some regions of the USA, higher BMI is correlated with the population density, for example, coastal southern counties of Texas, Mississippi, and Louisiana. Louisiana is the top state based on the prevalence rate (34.7%). As such, it is selected in Figure 2 and shows the spatial distribution of BMI by county.

3.2. Analysis

US sample size of 352,640 was used in the analysis with full records available for state, county, age, gender, and racial/ethnic-specific code. Distribution by sex was 42.48% male and 57.52% female. Adult population includes ages 18 and above. Five major racial/ethnic groups were analyzed: white (87.59%), black (9.36%), Asian (1.32%), Native Hawaiian or other Pacific Islanders (0.08%), and American Indian (1.64%). Alaska Native was not included in the sample due to missing data. Three main US regions were analyzed based on prevalence rates 20–25%, 25–30%, and 30–35% as reported by BRFSS (2012) appearing

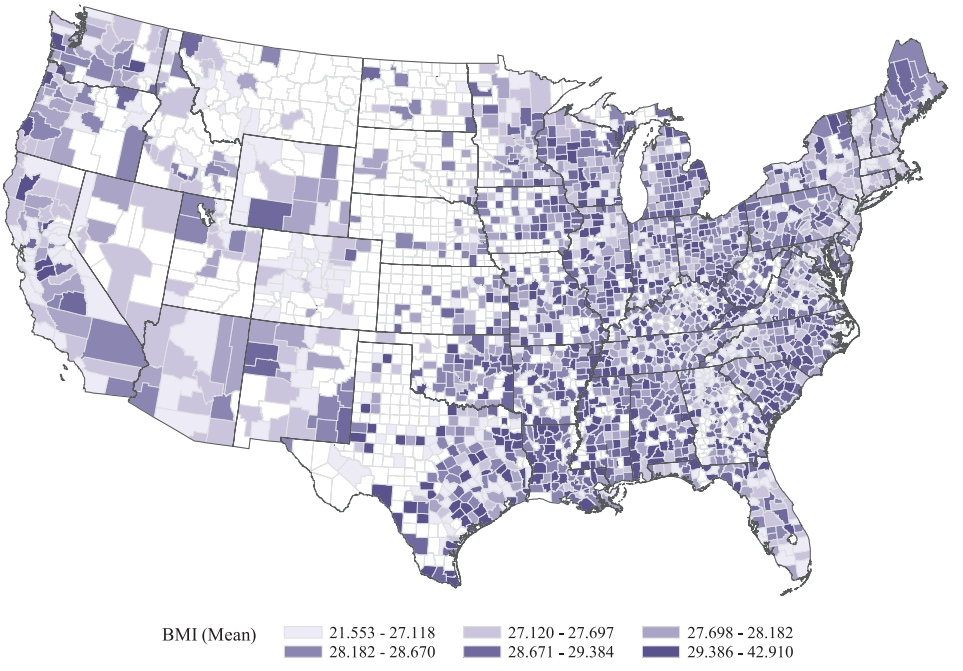

Figure 1. (Colour online) The US county map of average BMI values for the US adults.

at http://www.cdc.gov/obesity/data/adult.html. These regions are defined by state as follows:

- Prevalence rates 20–25%: CO, DC, MA, HI, NY, VT, MT, UT, NJ, and WY.
- Prevalence rates 25–30%: AZ, CA, CT, DE, FL, GA, ID, IL, KS, ME, MD, MN, MO, NE, NV, NH, NM, NC, ND, OR, PA, RI, SD, TX, VA, WA, and WI.
- Prevalence rates 30–35%: AL, AR, IN, IA, KY, LA, MI, MS, OH, OK, SC, TN, AND WV.

Due to a large sample size (> 2000), Jarque–Bera test, based on sample skewness and sample kurtoses, was used to test the normality assumption for BMI. The results of this test for the US BMI data has $\chi^2 = 309{,}433$ (df $= 2, p = 0.00$) rejecting the null hypothesis that the sample came from a normal distribution with expected skewness and kurtosis equal to zero. The conclusion is the same when running the Jarque-Bera test on the BMI data by prevalence region, gender, and race, indicating that the normality assumption is rejected. Another normality test, D'Agostino's K-squared test was run that measure departure from normality and it is based on transformation of the sample kurtosis and skewness. The result of this test for the US BMI has $K^2 = 92,947$ ($df = 2$, $p = 0.00$) leading to a conclusion that the distribution of BMI data is skewed. The conclusion is the same when running the D'Agostino's K-squared test on the BMI data by prevalence region, gender, and race.

Among all single parametric distributions tested in Table 1 and Figure 3, on the nationwide data, the log-normal has the highest log-likelihood, smallest AIC, BIC, and K–S values. This indicates Log-normal distribution is the best fit among all single component skewed models. However, it should be noted that Log-normal distribution is just

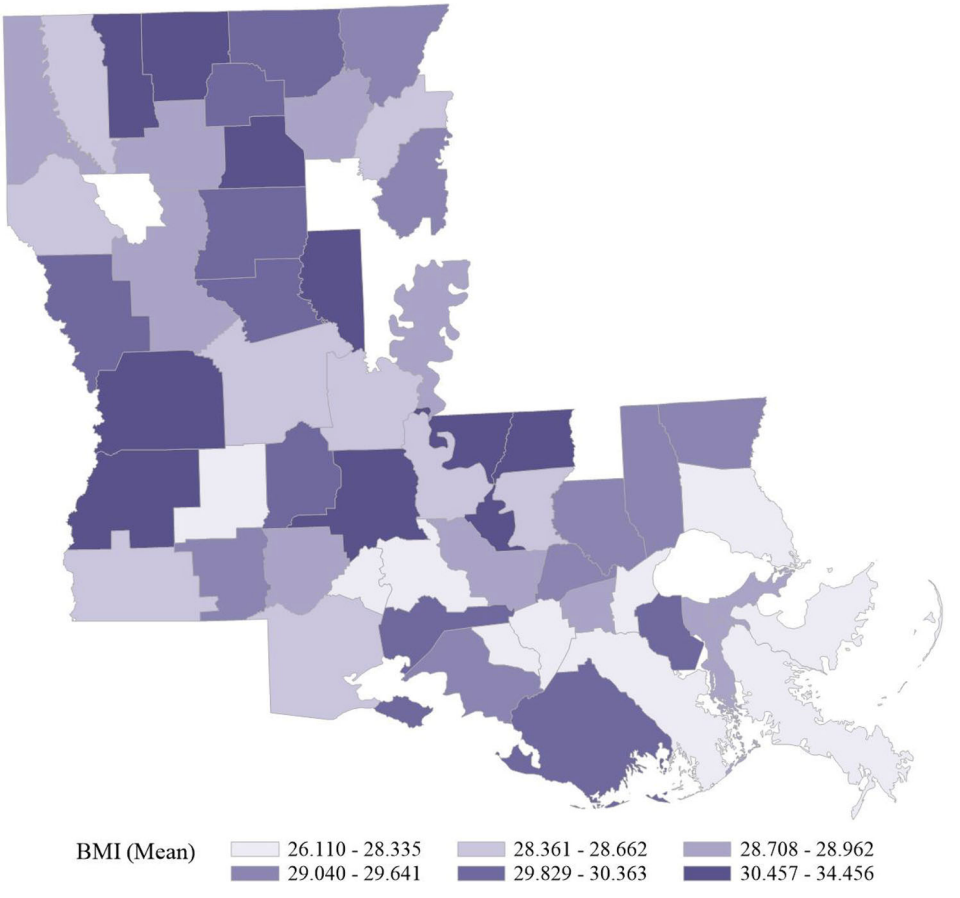

Figure 2. (Colour online) The Louisiana county map of average BMI values for the US adults.

marginally better than Inverse Gaussian (known as Wald). Thus, Inverse Gaussian distribution is also good fit for the BMI data. Parameters of the selected Log-normal distributions have mean of 3.31 and standard deviation of 0.20. However, the Log-normal model is inferior if compared to a 4-component Gaussian mixture model. The 4-component mixture shows higher log-likelihood and lower AIC, BIC, and K–S values compared to the log-normal model; therefore, represents the best fit overall.

Figure 4 compares the densities of the Log-normal and the 4-component mixture over the histogram of the data. It is obvious that by using the 4-component mixture, the fit is significantly improved in the peak and the right tail of the distribution compared to the Log-normal model, as shown in Figure 4. The centers of Gaussian subcomponents are $\mu_1 = 22.21$ ($\sigma_1 = 2.27$), $\mu_2 = 26.05$ ($\sigma_2 = 2.19$), $\mu_3 = 29.83$ ($\sigma_3 = 3.90$), and $\mu_4 = 35.47$ ($\sigma_4 = 8.45$). The mixing proportions by component are 0.23, 0.25, 0.37, and 0.15 reflecting the distribution of people within each subpopulation. About 29% of people in the USA have BMI above 30. The standard deviation increases for the third and fourth component to reflect better fit in the tail of the distribution. In addition to fitting the best model, these results illustrate that the finite mixture model can be used to determine an optimal number of subpopulations and their distributional properties within the overall population of BMI.

Table 1. Summary of distributions through 2012 BMI data for the US adults.

Distribution	Parameters	LogLik	AIC	BIC	K–S
Weibull	$\theta = 30.34, \tau = 4.29$	−1,164,926	2,329,856	2,329,878	0.108
Logistic	$\mu = 27.32, s = 3.23$	−1,121,339	2,242,682	2,242,704	0.049
Gamma	$\theta = 23.45, \alpha = 0.84$	−1,112,656	2,225,316	2,225,337	0.060
Inverse Gaussian	$\mu = 27.89, \theta = 656.27$	−1,105,752	2,211,508	2,211,479	0.048
Log-normal level	$\mu = 3.31, \sigma = 0.20$	−1,105,470	2,210,944	2,210,965	0.046
4-component	$\pi = (0.23, 0.25, 0.37, 0.15)$				
Gaussian mixture	$\mu = (22.21, 26.05, 29.83, 35.47)$	−1,100,256	2,200,534	2,200,653	0.014
	$\sigma = (2.27, 2.19, 3.90, 8.45)$				

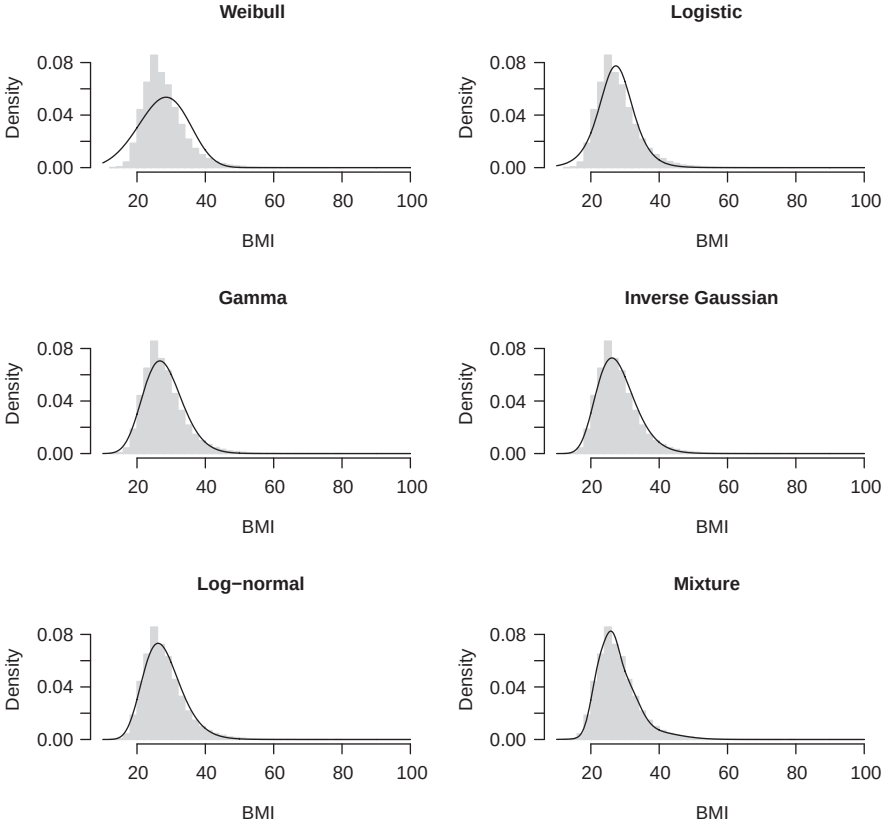

Figure 3. Different distributional fits through 2012 BMI data for the US adults.

The long thin tail of the distribution captures the observations with BMI $\geq$ 60 (1.1% of the data).

About 37% of the people form the subpopulation with the BMI mean value of 29.83 and standard deviation of 3.90. This largest subpopulation among the four subpopulations based on this model seems to be having different underlying causes of its obesity/overweight condition than the subpopulation whose mean BMI is 35.47 (15% of the total population) which is clearly reflecting the segment of obese people of all three grades: 1, 2 and 3. These two subpopulations should best be analyzed separately in multivariate regression framework to be able to more accurately identify the factors that underlie and

Table 2. Summary of distributions through 2012 BMI data by region based on prevalence rate.

Distribution	Parameters	LogLik	AIC	BIC	K–S
	Region with prevalence rate 20–25%				
Weibull	$\theta = 29.50, \tau = 4.42$	−238,803	477,610	477,729	0.112
Logistic	$\mu = 26.65, s = 3.02$	−229,164	458,332	458,350	0.051
Gamma	$\theta = 25.28, \alpha = 0.93$	−227,514	455,032	455,049	0.064
Inverse Gaussian	$\mu = 27.18, \theta = 691.57$	−226,008	452,020	452,038	0.050
Log-normal level	$\mu = 3.28, \sigma = 0.20$	−224,756	449,536	449,636	0.018
4-component	$\pi = (0.24, 0.25, 0.37, 0.14)$				
Gaussian mixture	$\mu = (21.94, 25.50, 29.09, 34.36)$ $\sigma = (2.16, 2.08, 3.74, 8.12)$	−224,756	449,536	449,636	0.018
	Region with prevalence rate 25–30%				
Weibull	$\theta = 30.25, \tau = 4.33$	−624,929	1,249,862	1,249,881	0.108
Logistic	$\mu = 27.29, s = 3.18$	−601,109	1,202,222	1,202,242	0.049
Gamma	$\theta = 24.01, \alpha = 0.86$	−596,601	1,193,206	1,193,227	0.059
Inverse Gaussian	$\mu = 27.83, \theta = 670.59$	−592,996	1,185,996	1,185,968	0.047
Log-normal level	$\mu = 3.30, \sigma = 0.20$	−592,841	1,185,686	1,185,706	0.046
4-component	$\pi = (0.24, 0.25, 0.37, 0.14)$				
Gaussian mixture	$\mu = (22.21, 26.07, 29.76, 35.27)$ $\sigma = (2.26, 2.17, 3.92, 8.39)$	−590,204	1,180,430	1,180,543	0.015
	Region with prevalence rate 30–35%				
Weibull	$\theta = 31.17, \tau = 4.18$	−299,559	599,122	599,244	0.103
Logistic	$\mu = 27.97, s = 3.44$	−289,488	578,980	578,999	0.048
Gamma	$\theta = 21.66, \alpha = 0.78$	−287,585	575,174	574,139	0.055
Inverse Gaussian	$\mu = 28.58, \theta = 619.25$	−285,328	570,660	570,530	0.042
Log-normal level	$\mu = 3.33, \sigma = 0.21$	−285,266	570,536	570,555	0.041
4-component	$\pi = (0.24, 0.24, 0.35, 0.17)$				
Gaussian mixture	$\mu = (22.51, 26.52, 30.49, 36.33)$ $\sigma = (2.43, 2.25, 3.91, 8.66)$	−284,119	568,260	568,364	0.013

correlate with corresponding BMIs. This is important in a sense that different causes call for different actions in terms of prevention, health care and policy. If one is to use BMI benchmark of 30, this large subpopulation is likely to be arbitrarily divided on those with BMI below or above 30 without accounting for true underlying causes for their condition. Than individuals with BMI of 27 and 33, for example, are likely to have more similarities in terms of causes of their condition then individuals with BMIs of 33 and 39, or 27 and 21 respectively. However, based on current classification of individuals on obese versus not obese which is based on the BMI benchmark of 30, this conclusion would not be possible.

Regional BMI data are analyzed based on the prevalence rate definition set by CDC. Summary of the results is displayed in Table 2. In case of all three regions based on prevalence rates 20–25%, 25–30%, and 30–35%, 4-component Gaussian mixtures is the best fitted model. The parameter estimates are fairly similar for the three groups and there is a visible increase in skewness as the prevalence rate increases. Mean and standard deviations of each component slightly increases across three groups. The highest means and standard deviations can be observed in the region with prevalence rate 30–35%.

Table 3 summarizes results for several models used to fit 2012 BMI data by gender. Gaussian mixture model represents the best fit. Female BMI can be described using a 4-component mixture with the following parameter estimates: $\mu_1 = 21.54$ ($\sigma_1 = 2.06$), $\mu_2 = 25.35$ ($\sigma_2 = 2.26$),$\mu_3 = 29.62$ ($\sigma_3 = 4.07$), $\mu_4 = 35.85$ ($\sigma_4 = 8.62$). These results indicate that the first subpopulation of about 24% of the total population is in normal range with no reasons to worry about potential overweight and obesity. However, studying this group can provide insight on behavioral, genetic, socio-economic or other relevant factors

Table 3. Summary of distributions through 2012 BMI data by gender.

Distribution	Parameters	LogLik	AIC	BIC	K–S
		Females			
Weibull	$\theta = 30.26, \tau = 4.01$	−682,355	1,364,714	1,364,845	0.113
Logistic	$\mu = 26.99, s = 3.50$	−661,061	1,322,126	1,322,147	0.062
Gamma	$\theta = 20.01, \alpha = 0.72$	−653,624	1,307,252	1,307,272	0.071
Inverse Gaussian	$\mu = 27.67, \theta = 556.59$	−648,856	1,297,716	1,297,578	0.057
Log-normal level	$\mu = 3.29, \sigma = 0.22$	−648,782	1,297,568	1,297,588	0.055
4-component	$\pi = (0.24, 0.24, 0.36, 0.16)$				
Gaussian mixture	$\mu = (21.54, 25.35, 29.62, 35.85)$ $\sigma = (2.06, 2.26, 4.07, 8.62)$	−644,447	1,288,916	1,289,029	0.020
		Males			
Weibull	$\theta = 30.39, \tau = 4.79$	−479,192	958,388	958,479	0.117
Logistic	$\mu = 27.72, s = 2.80$	−456,467	912,938	912,957	0.043
Gamma	$\theta = 30.76, \alpha = 1.09$	−454,771	909,546	909,566	0.054
Inverse Gaussian	$\mu = 28.18, \theta = 871.12$	−452,359	904,722	904,623	0.042
Log-normal level	$\mu = 3.32, \sigma = 0.18$	−452,181	904,366	904,386	0.041
3-component	$\pi = (0.39, 0.49, 0.12)$				
Gaussian mixture	$\mu = (25.25, 28.86, 34.59)$ $\sigma = (2.72, 4.19, 8.11)$	−449,423	898,862	898,942	0.024

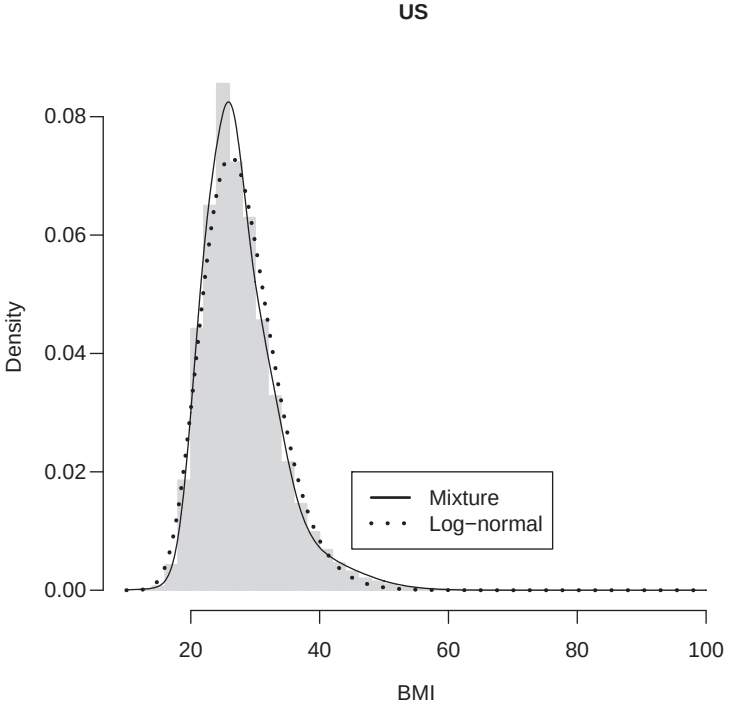

Figure 4. Log-normal vs. optimal Gaussian mixture model of the BMI of the US adults.

underlying their BMI status. Likewise, second group or subpopulation is of the same size, but individuals in this group are on the 'verge' of overweight. Hence, studying underlying factors of this group's BMI should be helpful in designing more effective preventative behavioral and health policy measures that would preclude its members from potentially

moving into the obese category. The third group, which is the largest of all female subpopulations containing 36% of the total females, is comprised of individuals who belong to either overweight or obese category. Studying underlying factors of BMI for this subpopulation should be helpful in developing medical treatments and creating policies that could remedy their current condition, and reverse or at least halt at its current level their BMI. These are likely to be different from the treatments and policies that would adequately address needs of the fourth subpopulation (16% of the total female population) that could be classified as severely obese, and some morbidly obese individuals since underlying causes for their BMI condition are likely differ.

Male BMI population can be best defined using 3-component Gaussian mixture with the following parameters:$\mu_1 = 25.25$ ($\sigma_1 = 2.72$), $\mu_2 = 28.86$ ($\sigma_2 = 4.19$), $\mu_3 = 34.58$ ($\sigma_3 = 8.11$). What is immediately obvious based on these results is that virtually all of male population is in danger of being overweight or obese. Alternatively, this result points to potentially serious deficiency of current classification of males on normal, overweight and obese based on BMI values being less than 25, between 25 and 30, and above 30, respectively. While influential previous literature points that prevalence of obesity, and especially extreme obesity, in females exceeds the prevalence of obesity in males [28], they fail to recognize that overall male population is more impacted by overweight and obesity (based on current definitions and classifications for both) unless there is recognition that current standards, and definitions for overweight and obesity based on them, are not applicable and are changed. For instance, it has been long recognized that men have more skeletal muscle mass than women and that these gender differences are greater in the upper body [20]. This greater skeletal muscle mass in men may lead to higher values of BMI for males of the same height as females. However, these higher BMI values are not necessarily an indication of overweight or obesity as current BMI based-standards would suggest.

The results of an analysis of distributions of BMI by racial/ethnic group were presented in Table 4. Since the assumption of normality is violated for all Jarque-Bera tests by race, normal distribution was eliminated for further considerations. For all racial/ethnic groups mixture model is selected as the best fit. Among single components, Log-normal and Inverse Gaussian are the best fits. The BMI for White population can be described by 4-component mixture with the following parameter estimates by component: $\mu_1 = 22.14$ ($\sigma_1 = 2.24$), $\mu_2 = 25.93$ ($\sigma_2 = 2.15$), $\mu_3 = 29.58$ ($\sigma_3 = 3.84$), and $\mu_4 = 35.07$ ($\sigma_4 = 8.28$). These results are very similar to the nation-wide results as the white population dominates the BMI data. About 28% of white people have BMI above 30.

The BMI for Blacks or African Americans show more shifts forward the right tail resulting in additional clustering and fitting a 5-component mixture model as the best model. The parameter estimates for this model are $\mu_1 = 22.93$ ($\sigma_1 = 2.47$), $\mu_2 = 26.07$ ($\sigma_2 = 1.80$), $\mu_3 = 29.49$ ($\sigma_3 = 2.32$), $\mu_4 = 33.18$ ($\sigma_4 = 4.53$), $\mu_5 = 38.25$ ($\sigma_5 = 9.59$). About 44% of the data are located in the 4th and 5th component of this mixture indicating that Blacks or African Americans are on average heavier than White people. The mean of single component models of Blacks or African Americans is shifted to the right compared to all other racial/ethnic groups indicating that average person in this group is heavier than those in all other groups. At least 43% of Blacks or African Americans have BMI above 30.

The BMI data for Asians is described best with a 2-component mixture with the estimated parameters: $\mu_1 = 23.53$ ($\sigma_1 = 3.01$),$\mu_2 = 28.78$ ($\sigma_2 = 5.56$). The mean of BMI for a single component model for Asians is significantly lower compared to other racial/ethnic

Table 4. Summary of distributions through 2012 BMI data by race.

Distribution	Parameters	LogLik	AIC	BIC	K–S
		White			
Weibull	$\theta = 30.07, \tau = 4.36$	−1,013,321	2,026,646	2,026,782	0.108
Logistic	$\mu = 27.14, s = 3.14$	−974,504	1,949,012	1,949,034	0.048
Gamma	$\theta = 24.32, \alpha = 0.88$	−967,275	1,934,554	1,934,575	0.059
Inverse Gaussian	$\mu = 27.68, \theta = 675.01$	−961,411	1,922,826	1,922,682	0.047
Log-normal level	$\mu = 3.30, \sigma = 0.20$	−961,163	1,922,330	1,922,352	0.046
4-component	$\pi = (0.24, 0.25, 0.37, 0.14)$				
Gaussian mixture	$\mu = (22.14, 25.93, 29.58, 35.07)$ $\sigma = (2.24, 2.15, 3.84, 8.28)$	−956,703	1,913,428	1,913,546	0.015
		Blacks or African Americans			
Weibull	$\theta = 32.82, \tau = 4.07$	−113,158	226,320	226,461	0.100
Logistic	$\mu = 29.38, s = 3.77$	−109,822	219,648	219,666	0.049
Gamma	$\theta = 20.08, \alpha = 0.67$	−108,811	217,626	217,642	0.054
Inverse Gaussian	$\mu = 30.03, \theta = 600.59$	−108,197	216,398	216,248	0.040
Log-normal level	$\mu = 3.38, \sigma = 0.22$	−108,177	216,358	216,374	0.039
5-component	$\pi = (0.21, 0.15, 0.19, 0.28, 0.17)$				
Gaussian mixture	$\mu = (22.93, 26.07, 29.49, 33.18, 38.25)$ $\sigma = (2.47, 1.80, 2.32, 4.53, 9.59)$	−107,834	215,696	215,814	0.013
		Asians			
Weibull	$\theta = 26.72, \tau = 5.22$	−13,904	27,812	27,824	0.109
Logistic	$\mu = 24.48, s = 2.38$	−13,302	26,608	26,622	0.042
Gamma	$\theta = 33.81, \alpha = 1.36$	−13,238	26,480	26,492	0.057
Inverse Gaussian	$\mu = 24.84, \theta = 841.79$	−13,175	26,354	26,333	0.047
Log-normal level	$\mu = 3.19, \sigma = 0.17$	−13,172	26,348	26,361	0.046
2-component	$\pi = (0.75, 0.25)$				
Gaussian mixture	$\mu = (23.53, 28.78)$ $\sigma = (3.01, 5.56)$	−13,129	26,268	26,301	0.019
		Native Hawaiian or other Pacific Islander			
Weibull	$\theta = 30.33, \tau = 4.63$	−957	1918	1943	0.102
Logistic	$\mu = 27.39, s = 3.22$	−932	1868	1875	0.066
Gamma	$\theta = 24.45, \alpha = 0.88$	−922	1848	1855	0.072
Log-normal	$\mu = 3.31, \sigma = 0.20$	−918	1840	1846	0.059
Inverse Gaussian	$\mu = 27.92, \theta = 681.18$	−917	1838	1823	0.060
2-component	$\pi = (0.54, 0.46)$				
Gaussian mixture	$\mu = (24.85, 31.46)$ $\sigma = (3.09, 6.29)$	−916	1842	1861	0.034
		American Indians			
Weibull	$\theta = 31.83, \tau = 4.25$	−19,483	38,970	39,035	0.095
Logistic	$\mu = 28.71, s = 3.49$	−18,852	37,708	37,720	0.043
Gamma	$\theta = 22.09, \alpha = 0.76$	−18,717	37,438	37,451	0.050
Inverse Gaussian	$\mu = 29.24, \theta = 640.01$	−18,634	37,268	37,199	0.036
Log-normal level	$\mu = 3.35, \sigma = 0.21$	−18,629	37,262	37,276	0.035
3-component	$\pi = (0.37, 0.49, 0.14)$				
Gaussian mixture	$\mu = (25.26, 30.06, 37.06)$ $\sigma = (3.43, 4.91, 8.98)$	−18,598	37,212	37,265	0.021

groups. The second component in this mixture includes 25% of the population with mean BMI of 28.78 indicating that less than 25% of Asians have BMI 30 and above. In fact, about 11% of the Asians have BMI above 30.

The best 3-component mixture for American Indians BMI has the following estimated parameters: $\mu_1 = 25.26$ ($\sigma_1 = 3.43$), $\mu_2 = 30.06$ ($\sigma_2 = 4.91$), and $\mu_3 = 37.06$ ($\sigma_3 = 8.98$). Results for this group could best indicate the advantage of using the best (true) underlying distribution, that is, the Gaussian mixture, rather than the best single component distribution model. Mean of the best single component parametric model for

American Indians is slightly lower than the mean for Blacks or African American indicating that on average about 38% of people in this group has BMI above 30. What these results are not telling us and what the results from the Gaussian mixture are telling us is that virtually all of American Indian population is affected by overweight and obesity, while there is still a sizeable subpopulation of the Black or African Americans who are in normal range, based on current definitions. While extreme obesity is more represented among the Black or African Americans, the overweight and obesity public health problem among American Indians is actually most acute.

Native Hawaiian and Other Pacific Islanders are the smallest ethnic/racial group. Due to the sample size, only 2-component mixture is selected as the best model with $\mu_1 = 24.85$ ($\sigma_1 = 3.09$), $\mu_2 = 31.46$ ($\sigma_2 = 6.29$). For this group, a single component Inverse Gaussian model is slightly better than the Log-normal model. About 30% of Native Hawaiians have BMI above 30.

4. Conclusion

True distribution of BMI, as an accepted or standard measure of obesity, is needed not only to describe the variable and phenomenon properly, but to help analyze its underlying causes in more complex, multivariate framework. Based on BRFSS survey data, the BMI of US adult population is best described by a 4-component Gaussian mixture model. This model has the highest value of the log-likelihood function among all right skewed models analyzed. A 4-component mixture model with different parameters also applies to population of white people, females, and all regions based on the prevalence rate division. These four components or subpopulations reflect heterogeneous characteristics of the population. A 5-component mixture is the best model for fitting BMI data for Black or African American population. The BMI for males and American Indians can be described using 3-component mixture. Finally, 2-component mixture fits BMI data for Asians and Native Hawaiians. As the general population of BMI tends to shift over time in the upper tail, the optimal number of components as well as the parameters of each component should be tested with availability of new BMI data.

Describing prevalence of obesity based on true probability distribution of the BMI enables us to analyze and determine underlying causes of BMI status for each subpopulation, and in turn design more suitable and effective medical treatments or food and health policy measures. The immediate convenience of 'one size fits all' medical treatments and food and health policies is obvious for the health, nutrition or fitness practitioners and policy makers alike. Yet the long-term inefficiency of measures that do not account for diversity of underlying causes of overweight and obesity for different subpopulations is likely to far exceed these short-term convenience-based advantages. Hence, it may be time to reconsider the guidelines for how to classify population on obese, overweight, or normal based on BMI since the current standards are fairly dated and do not account for the diverse underlying factors among different subpopulations that lead towards overweight and obesity. Proposed method may add some complexity to the process, but it increases probability of more accurate representation of prevalence of overweight and obesity, and potentially enables us to identify more accurately the underlying reasons for it for various subpopulations. This approach would inevitably lead to more diverse, and, hopefully,

more successful treatment of this premier public health problem by both medical practitioners and health and food policy makers. Finally, findings of the study are limited to BRFSS survey data. In order to make this claim more general, it would be useful to pursue future research and simulate data from different distributions base on different surveys (NHANES, NHIS, BRFSS) or compare analyses from these three data bases.

Disclosure statement

No potential conflict of interest was reported by the authors.

References

[1] H. Akaike, *A new look at the statistical modelling identification*, IEEE Trans. Automat. Control 19 (1974), pp. 716–723.
[2] F.D. Atem, R.K. Sharma, and S.J. Anderson, *Fitting bivariate multilevel models to assess long-term changes in body mass index and cigarette smoking*, J. Appl. Stat. 38 (2011), pp. 1819–1831.
[3] V. Bewick, L. Cheek, and J. Ball, *Statistics review 14: Logistic regression*, Crit. Care 9 (2005), pp. 112–118.
[4] J.M. Bland and D.G. Altman, *Transforming data*, BMJ: British Medical Journal 312 (1996), p. 770.
[5] G.A. Bray, C. Bouchard, T.S. Church, W.T. Cefalu, F.L. Greenway, A.K. Gupta, L.M. Kaplan, E. Ravussin, S.R. Smith, and D.H. Ryan, *Is it time to change the way we report and discuss weight loss?* Obesity 17 (2009), pp. 619–621.
[6] CDC, *Morbidity and Mortality Weekly. Obesity – United States, 1999-2000*. Available at www.cdc.gov/mmwr/preview/mmwrhtml/su6203a20.htm.
[7] Centers for Disease Control and Prevention, *Prevalence of Self-reported Obesity Among US Adults BRFSS, 2012 by State*. Available at http://www.cdc.gov/obesity/data/adult.html, accessibility verified August 6, 2014 (source of the data).
[8] Centers for Disease Control and Prevention (CDC), *Factor Surveillance System Survey Data. Atlanta, Georgia: U.S*, Department of Health and Human Services, Centers for Disease Control and Prevention, 2012.
[9] S.-Y. Chou, M. Grossman, and H. Saffer, *An economic analysis of adult obesity: Results from the behavioral risk factor surveillance system*, J. Health Econ. 23 (2004), pp. 565–587.
[10] A.D.N. Drewnowski, *The economics of obesity: Dietary energy density and energy cost*, Am. J. Clin. Nutr. 82 (2005), pp. 265S–273S.
[11] Expert Panel on the Identification, Evaluation, and Treatment of Overweight in Adults, *Clinical guidelines on the identification, evaluation, and treatment of overweight and obesity in adults: Executive summary*, Am. J. Clin Nutr. 68 (1998), pp. 899–917. Available at http://ajcn.nutrition.org/content/68/4/899.full.pdf, accessibility verified August 18, 2014.
[12] K.M. Flegal, M.D. Carroll, R.J. Kuczmarski, and C.L. Johnson, *Overweight and obesity in the United States: Prevalence and trends, 1960-1994*, Int. J. Obes. 22 (1998), pp. 39–47.
[13] K.M. Flegal, M.D. Carroll, B.K. Kit, and C.L. Ogden, *Prevalence of obesity and trends in the distribution of body mass index among US adults, 1999-2010*, JAMA 307 (2012), pp. 491–497.
[14] K.M. Flegal, M.D. Carroll, C.L. Ogden, and L.R. Curtin, *Prevalence and trends in obesity among US adults, 1999-2008*, JAMA 303 (2010), pp. 235–241.
[15] K.M. Flegal, M.D. Carroll, C.L. Ogden, and C.L. Johnson, *Prevalence and trends in obesity among US adults, 1999-2000*, JAMA 288 (2002), pp. 1723–1727.
[16] K.M. Flegal and R.P. Troiano, *Changes in the distribution of body mass index of adults and children in the US population*, Int. J. Obes. 24 (2000), pp. 807–818.
[17] R.M. Groves and R.L. Kahn, *Surveys by Telephone: A National Comparison With Personal Interviews*, Academic Press Inc, New York, 1979.
[18] R.M. Groves, P.P. Biemer, L.E. Lyberg, J.T. Massey, W.L. Nicholls, and J. Waksberg (eds.), *Telephone Survey Methodology*, John Wiley & Sons Inc, New York, 1988.

[19] D.L. Hartl and A.G. Clark, *Principles of Population Genetics*, Sinauer Associates, Inc, Sunderland (MA), 1997.
[20] I. Janssen, S.B. Heymsfield, Z.M. Wang, and R. Ross, *Skeletal muscle mass and distribution in 468 men and women aged 18–88 yr*, J. Appl. Physicol. 89 (2000), pp. 81–88.
[21] A.P. MacKay, L.A. Fingerhut, and C.R. Duran, *Health, United States, With Adolescent Chart Book*, National Center for Health Statistics, Hyattsville, MD, 2000, DHHS publication 2000:00–1232.
[22] G. Marsaglia, W.W. Tsang, and J. Wang, *Evaluating Kolmogorov's distribution*, J. Stat. Softw. 8 (2003), pp. 1–4.
[23] G. McLachlan and D. Peel, *Gaussian Mixture Models*, Wiley & Sons, Hobuken, NJ, 2000.
[24] D. Miljkovic and W. Nganje, *Economic factors affecting the increase in obesity in the United States: A model of myopic addictive behavior in food consumption*, Agric. Econ. 38 (2008), pp. 375–384.
[25] D. Miljkovic, W. Nganje, and H. de Chastenet, *Economic factors affecting the increase in obesity in the United States: The differential response to price*, Food Policy 33 (2008), pp. 48–60.
[26] A.H. Mokdad, M.K. Serdula, W.H. Dietz, B.A. Bowman, J.S. Marks, and J.P. Koplan, *The continuing epidemic of obesity in the United States*, JAMA 284 (2000), pp. 1650–1651.
[27] D.E. Nelson, E. Powell-Griner, M. Town, and M.G. Kovar, *A comparison of national estimates from the national health interview survey and the behavioral risk factor surveillance system*, Am. J. Public Health 93 (2003), pp. 1335–1341.
[28] C.L. Ogden, M.D. Carroll, L.R. Curtin, M.A. McDowell, C.J. Tabak, and K.M. Flegal, *Prevalence of overweight and obesity in the United States, 1999-2004*, JAMA 295 (2006), pp. 1549–1555.
[29] A.D. Penman and W.D. Johnson, *The changing shape of the body mass index distribution curve in the population: Implications for public health policy to reduce the prevalence of adult obesity*, J. PVC Prev Chronic Dis. 3 (2006), pp. 1–4.
[30] G. Schwarz, *Estimating the dimension of model*, Ann. Statist. 6 (1978), pp. 461–464.
[31] S.R. Smith and D.H. Ryan, *Is it time to change the way we report and discuss weight loss?* Obesity 17 (2009), pp. 619–621.
[32] S.T. Stewart, D.M. Cutler, and A.B. Rosen, *Forecasting the effects of obesity and smoking on US life expectancy*, N. Engl. J. Med. 361 (2009), pp. 2252–2260.
[33] The R project for Statistical Computing, Available at website: http://www.r-project.org/.
[34] U.S. Department of Health and Human Services (HHS), *Healthy People 2010*, Washington, DC, 2000. Available at http://www.healthypeople.gov.
[35] Y. Wang, M.A. Beydoun, L. Liang, B. Caballero, and S.K. Kumanyika, *Will all Americans become overweight or obese? estimating the progression and cost of the US obesity epidemic*, Obesity 16 (2008), pp. 2323–2330.
[36] S. Yun, B.-P. Zhu, W. Black, and R.C. Brownson, *A comparison of national estimates of obesity prevalence from the behavioral risk factor surveillance system and the national health and nutrition examination survey*, Int. J. Obes. 30 (2006), pp. 164–170.