Skewed and Flexible Skewed Distributions: A Modern Look at the Distribution of BMI

Thao Tran, Cara Wiskow, and Mohammad Abdus Aziz*
Department of Mathematics, University of Wisconsin–Eau Claire, Eau Claire, WI

Students: tranp@uwec.edu, wiskowca@uwec.edu*

Mentor: azizm@uwec.edu

ABSTRACT
The purpose of this study is to find distributions that best model body mass index (BMI) data. BMI has become a standard health indicator and numerous studies have been done to examine the distribution of BMI. Due to the skew and bimodal nature, we focus on modeling BMI with flexible skewed distributions. The distributions are fitted to University of Wisconsin–Eau Claire (UWEC) BMI data and to data obtained from National Health and Nutrition Survey (NHANES). The model parameters are obtained using maximum likelihood estimation method. We compare flexible models to more conventional distributions, such as skew-normal and skew-t distributions, using AIC and BIC and Kolmogorov-Smirnov (K-S) goodness-of-fit test. Our results indicate that the skew-t and Alpha-Skew-Laplace distributions are reasonably competitive when describing unimodal BMI data whereas Alpha-Skew-Laplace, finite mixture of scale mixture of skew-normal and skew-t distributions are better alternatives to both unimodal and bimodal conventional distributions. The results we obtained are useful because we believe the models discussed in our study will offer a framework for testing features such as bimodality, asymmetry, and robustness of the BMI data, thus providing a more detailed and accurate understanding of the distribution of BMI.

KEYWORDS
Body Mass Index; Skew-normal distribution; Skew-t distribution; Flexible skewed distributions; Mixture distributions; Scale mixture of skew-normal distribution; K-S test

INTRODUCTION

Obesity has been reaching epidemic proportions in the United States. The rise has implications both on health and health care costs. Body mass index (BMI; kg/m^2) in the overweight (25 to 29.9) or obese category (30 or above) has been linked to cancer, hypertension, heart failure, cardiovascular disease, diabetes, stroke, and more. Because of this, obesity places a burden on the health care system, raising costs for the public. Given these negative impacts, governments and organizations have been actively trying to reduce obesity. In order to better assess obesity risk and address its prevalence, a better understanding of the distribution of BMI is imperative.

BACKGROUND

Since obesity became a major public health concern, many distributions have been applied to BMI data in an effort to find the best way to describe it. Multiple probability distributions have been fitted to Australian athletes' BMI data. Examples for the univariate case include Ma et al.'s generalized skew-normal (GSN) distribution,[1] Canale's extended skew-normal (ESN),[2] and Olivares-Pacheco et al.'s epsilon-skew-slash (ESS) and epsilon-skew-normal (ESN)[3]. For multivariate analysis of the same data set, Tan et al. considered the skew-slash t (SSLT) and skew-slash-normal (SSLN)[4] distributions and Arslan[5] applied generalized hyperbolic (GH) and variance-mean mixture of skew-normal (SN) distributions. When analyzing potential effectiveness of a tax on sweetened beverages in South Africa, Manyema et al.[6] applied log-normal and gamma distributions to describe the BMI distribution in their data set. For a US data set, Miljkovic et al.[7] found the k-component

Gaussian mixture distribution best fit their BMI data when compared to the log-normal, Weibull, logistic, inverse Gaussian, and gamma distributions.

While plenty of work has been done to examine unimodal BMI distributions, little has been done concerning bimodal BMI data. Mixture distributions provide a way to model bimodal data well. Lin et al.[8] investigated the potential for describing a bimodal BMI data set using mixture of normal, mixture of skew-normal, mixture of student's t, and mixture of skew-t distributions.

In this paper, we explore the abilities of various distributions to model both UWEC BMI and NHANES BMI data. Recent developments have been made in distributions which can account for either unimodality or bimodality of skewed data. They comprise alpha skew-normal, a new family of distributions, introduced by Elal-Olivero, that is flexible to both unimodal and bimodal data.[9] In 2012, Harandi et al.[10] presented a new class of skew distributions of Elal-Olivero's family using Laplace distribution, known as Alpha-Skew-Laplace distribution. This distribution can fit unimodal and bimodal shapes with increasing and u-shaped hazard functions for the truncated case at the origin. In 2014, the simple approach was used with logistic functions by Hazarika et al.,[11] who proposed a new distribution called alpha-skew-logistic distribution. This distribution can especially model the data given either positive or negative skewness. Another recently proposed model is the bimodal skew-symmetric normal distribution by Hassan et al.,[12] which can resolve problems of asymmetry, platykurtic/leptokurtic data (exhibit excess negative/positive kurtosis), and different types of bimodality. Kollu et al. also introduced three new mixture models, including Weibull-log-normal, GEV-log-normal, and Weibull-GEV.[13] We apply some of these newly introduced distributions as well as other mixture distributions to two BMI data sets in order to see if they provide a more accurate description of BMI distributions.

Significance

In recent years, there has been considerable interest in skewed distributions,[14] and considering the skew and bimodal features of some of the medical data, it is really important that appropriate distributions be fitted to these kinds of data. The use of statistical procedures is inappropriate if the actual distribution differs from the assumed type. This study mainly focuses on a statistical practice of fitting skew-symmetric distributions to medical data with bimodal characteristics while still considering unimodal and mixture distributions.

The most commonly used techniques for modeling bimodality involve using the mixture distributions. However, the proposed models created computational implementation difficulties. Many authors have also proposed different versions of the bimodal normal distribution to replace mixture distributions but because they fail to take into account the asymmetry, these studies did not materialize in the real world of statistics.

Many medical data sets, including that of BMI, are bimodal, asymmetric and platykurtic, or leptokurtic. We believe that the models we consider in this study will offer a framework for testing these features of the data at hand and overcome some of the complexities of the existing models. If these features are found to be significant, the proposed distribution will provide the user with a parsimonious model that will fit the data adequately. Thus, the user can test symmetry, excess kurtosis, and bimodality in order to adjust the values of model parameters accordingly to ensure a good fit.

The uniqueness and skewness of the data make it difficult to model perfectly with many known distributions, but the flexibility of the skew-symmetric distribution offers an increasingly insightful perspective that could offer a solution in dealing with bimodality of data in the field of medicine.

METHODS AND PROCEDURES

Flexible Skew-Symmetric Distributions

Unimodal Skew-Symmetric Distributions

1. Skew-normal distribution

   A skew-normal distribution has the following probability density function (pdf)

$$f(y;\alpha) = 2\phi(y)\Phi(\alpha y),\ y \in R \qquad \textbf{Equation 1.}$$

   where ϕ represents the density distribution of $N(0,1)$, Φ represents the cumulative distribution of $N(0,1)$, and α represents the shape parameter. We use the notation $SN(\alpha)$ to denote a skew-normal random variable.
   The linear transformation $X = \mu + \alpha Y$ with $\mu \in R$ and $\sigma > 0$ has the density of

$$f(x;\mu,\sigma,\alpha) = \frac{2}{\sigma}\phi\Big(\frac{x-\mu}{\sigma}\Big)\Phi\Big(\alpha\frac{x-\mu}{\sigma}\Big),\ x \in R. \qquad \textbf{Equation 2.}$$

   Then $X \sim SN(\mu,\sigma,\alpha)$, which reduces to the standard skew-normal distribution when $X \sim SN(0,1,\alpha)$. If α is set to 0, the distributions become the pdf of a standard normal distribution.
   The skew-normal cumulative distribution function is given by the following equation:

$$\Phi(z,\lambda) = 2\int_{-\infty}^{z}\int_{-\infty}^{\lambda t}\phi(t)\phi(u)\,dudt. \qquad \textbf{Equation 3.}$$

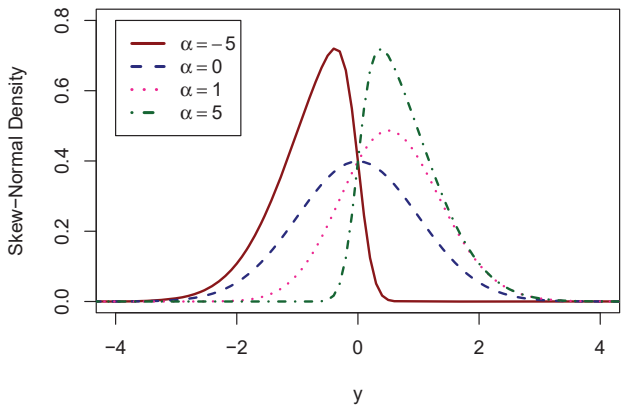

Figure 1. Density plot of skew-normal distribution for some selected values of α.

2. Skew-t distribution

   Let Z be a standard skew-normal random variable and W be a variable with $\chi^2(\nu)$ distribution. Suppose Z and W are independent. Define

$$Y = \frac{Z}{\sqrt{\frac{W}{\nu}}}. \qquad \textbf{Equation 4.}$$

   Then the linear transformation $X = \mu + \sigma Y$ has a skew-t distribution with parameters μ, σ, α, and ν and we introduce the notation $ST(\mu,\sigma,\alpha,\nu)$ to denote the skew-t random variable X.

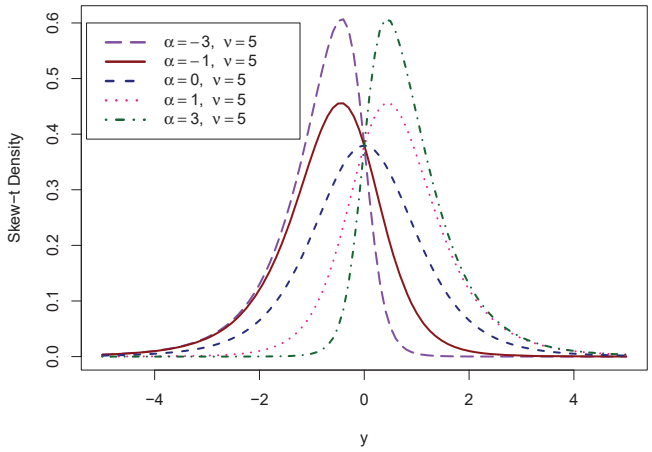

Figure 2. Density plot of skew-t distribution for some selected values of α and ν.

3. Generalized extreme value distribution

   The random variable X is said to have generalized extreme value (GEV) distribution when X has the following density function

$$e(x;\zeta,\delta,l) = \left(\frac{l}{\delta}\right)\left(1+\frac{\zeta(x-1)}{\delta}\right)^{-\frac{1}{\zeta}-1} e^{-\left(1+\frac{\zeta(x-1)}{\delta}\right)^{\frac{1}{\zeta}}} \qquad \textbf{Equation 5.}$$

   where $\zeta \neq 0$. The cumulative distribution of X is given by

$$E(x;\zeta,\delta,l) = e^{-\left(1+\frac{\zeta(x-1)}{\delta}\right)^{\frac{-1}{\delta}}}. \qquad \textbf{Equation 6.}$$

We do not provide the mathematical description of traditional distributions such as Gamma, Weibull, and log-normal. Their pdfs can be found in any standard book on distributions, for example, Balakrishnan and Johnson.[15]

Bimodal Skew-Symmetric Distributions

1. Alpha-skew-normal distribution

   A continuous random variable Y has an alpha-skew-normal distribution with a probability density function

$$f(y;\alpha) = \frac{(1-\alpha y)^2+1}{2+\alpha^2}\phi(y),\ y \in R \qquad \textbf{Equation 7.}$$

   where α represents the shape parameter. We denote this density as $Y \sim ASN(\alpha)$. If we adjust the pdf to include location and scale parameters the density becomes

$$f(y;\mu,\sigma,\alpha) = \frac{\left[1-\alpha\left(\frac{y-\mu}{\sigma}\right)\right]^2+1}{\sigma\left(2+\alpha^2\right)}\phi\left(\frac{y-\mu}{\sigma}\right),\ y \in R. \qquad \textbf{Equation 8.}$$

   Alpha-skew-normal has cumulative distribution function

$$F(y) = \Phi(y) + \alpha\left(\frac{2-\alpha y}{2+\alpha^2}\right)\phi(y). \qquad \textbf{Equation 9.}$$

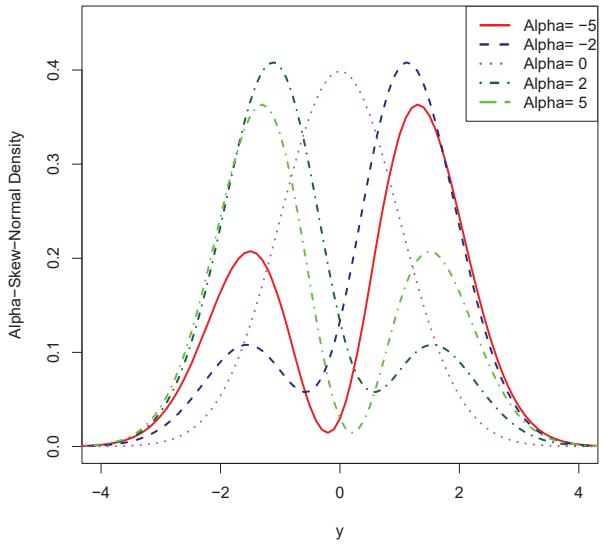

Figure 3. Density plot of alpha-skew-normal distribution for some selected values of α.

2. Alpha-skew-logistic distribution

Let Y be an alpha-skew-logistic random variable with parameter α then Y has density function

$$f(y;\alpha) = \frac{3((1-\alpha y)^2+1)e^{-y}}{(6+(\alpha^2\pi^2))(1+e^{-y})^2}. \qquad \textbf{Equation 10.}$$

We denoted it by $y \sim ASLG(\alpha)$. If α equals 0, we get the standard logistic distribution given by

$$f_y(y) = \frac{e^{-y}}{(1+e^{-y})^2}, \quad y \in R. \qquad \textbf{Equation 11.}$$

The Alpha-skew-logistic cumulative distribution function is given by

$$F(y;\alpha) = \frac{3}{6+\alpha^2\pi^2}\left(\frac{(1-\alpha y)^2+1}{1+e^{-y}} + 2\alpha(1-\alpha y)\log(1+e^{y}) - 2\alpha^2 Li_2(-e^{y})\right). \qquad \textbf{Equation 12.}$$

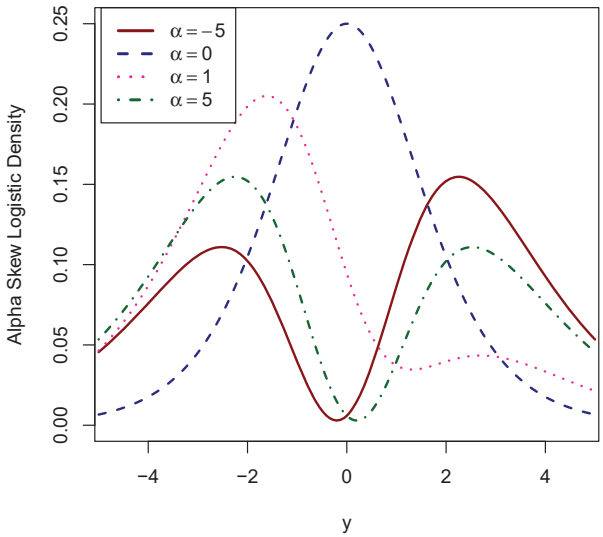

Figure 4. Density plot of Alpha-skew-logistic distribution with chosen α values -5, 0, 1, 5.

3. Alpha-Skew-Laplace distribution

A continuous random variable Y is said to follow an Alpha-Skew-Laplace distribution if its pdf has the form

$$f(y) = \frac{(1-\alpha y)^2+1}{4(1+\alpha^2)} e^{-|y|},\ y \in R \qquad \textbf{Equation 13.}$$

where α represents the shape parameter. An Alpha-Skew-Laplace random variable is denoted by $ASLP(\alpha)$.

Suppose $Y \sim ASLP(\alpha)$. Then ASLP density of location and scale is defined as the distribution of $X = \mu + \sigma Y$ for $\mu \in R$ and $\sigma > 0$. The corresponding density function is given by

$$f(x) = \frac{(1-\alpha\frac{x-\mu}{\sigma})^2+1}{4\sigma(1+\alpha^2)} e^{-\frac{|x-\mu|}{\sigma}},\ x \in R \qquad \textbf{Equation 14.}$$

where $\theta = (\mu, \sigma, \alpha)$.

Alpha-Skew-Laplace cumulative distribution is given by the following.

$$F(t;\alpha) = \frac{1+(1-\alpha t)^2}{4(1+\alpha^2)} e^{t} + \frac{\alpha(1+\alpha(1-t))}{2(1+\alpha^2)} e^{t},\ t < 0 \qquad \textbf{Equation 15.}$$

$$F(t;\alpha) = 1 - \frac{1+(1-\alpha t)^2}{4(1+\alpha^2)} e^{-t} + \frac{\alpha(1+\alpha(1-t))}{2(1+\alpha^2)} e^{-t},\ t \geq 0$$

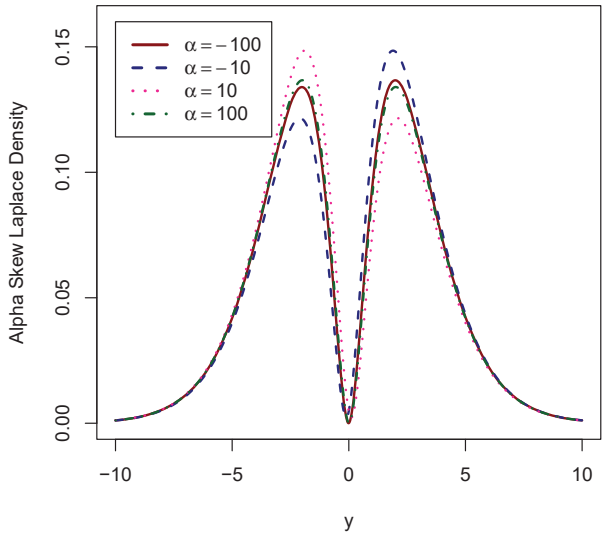

Figure 5. Density plot of Alpha-Skew-Laplace Distribution with chosen α values -5, 0, 1, 5.

4. Bimodal skew-symmetric normal distribution

   The random variable Y is said to have a bimodal skew-symmetric normal distribution when Y has the density distribution as follows

$$\Psi(y) = \Phi(y) - \frac{y + \mu - 2\beta}{1 + 2\psi(\delta + (\beta - \mu)^2))}\phi(y) \qquad \textbf{Equation 16.}$$

   where $\mu, \beta \in R$ represents the location parameter, $\psi > 0$ represents the shape parameter, and $\theta > 0$ represents the bimodality parameter.

Mixture Distributions

In this section we consider several conventional mixture distributions and recently developed scale mixture of skew-normal and skew-t distributions. In the mixture model context the density of x is expressed as a mixture of P parametric densities such that

$$f(x, \psi) = \sum_{i=1}^{p} \pi_i f(x; \theta_i). \qquad \textbf{Equation 17.}$$

The $\pi_i \geq 0, i = 1, 2, \ldots, p$ with $\sum_{i=1}^{p} \pi_i = 1$ are called mixing weights of the ith component of the mixture, which is characterized by parameter θ_i, and $\psi = (\pi_1, \pi_2, \ldots, \pi_{p-1}, \theta_1, \theta_2, \ldots, \theta_p)$ denotes the vector of parameters of the model.

1. Finite mixture of scale mixture of skew-normal distribution

   Suppose $Z \sim SN(0, \sigma^2, \alpha)$ and U be a positive random variable, independent of Z, with distribution function $H(u; \nu)$. Then the random variable $Y = \mu + U^{-1}Z$, where $\mu \in R$ is a location parameter, is said to follow a scale mixture of skew-normal (SMSN) distribution if its pdf is given by

$$f(y) = \int_0^\infty \phi\left(y; \mu, \sigma^2 u^{-1}\right) \Phi\left(u^{1/2}\alpha\left(\frac{y - \mu}{\sigma}\right)\right) dH(u). \qquad \textbf{Equation 18.}$$

   In the definition $H(.; \nu)$ is known as the *mixing scale distribution* and for each choice of this we get different members of the family such as normal, skew-normal, or student-t. A finite mixture of SMSN distributions[7] model is a density

defined as in Equation 17 where the ith component of the mixture is a SMSN density with parameters μ_i, σ_i^2, α_i, and ν_i. For simplicity we assume $\nu_1 = \nu_2 = \ldots = \nu$.

2. Two-component mixture Weibull distribution

   The two-component mixture Weibull distribution has five parameters and its probability distribution function is given as follows:

$$ff(x; k_1, c_1, k_2, c_2, w) = wf(x; k_1, c_1) + (1-w)f(x; k_2, c_2). \quad \textbf{Equation 19.}$$

   Its cumulative distribution function is given by

$$FF(x; k_1, c_1, k_2, c_2, w) = wF(x; k_1, c_1) + (1-w)F(x; k_2, c_2). \quad \textbf{Equation 20.}$$

3. Mixture gamma and Weibull distribution

   A random variable X is said to have mixture gamma and Weibull distribution when X has the following probability distribution:

$$h(x; \alpha, \beta, k, c, w) = wg(x, \alpha, \beta) + (1-w)f(x; k, c). \quad \textbf{Equation 21.}$$

   The mixture gamma and Weibull cumulative distribution function is given by

$$H(x; \alpha, \beta, k, c, w) = wG(x, \alpha, \beta) + (1-w)F(x; k, c). \quad \textbf{Equation 22.}$$

4. Mixture normal distribution

   A single-truncated normal probability distribution function is given by

$$q(x; \mu, \sigma) = \frac{1}{I(\mu, \sigma)\sigma\sqrt{2\pi}} e^{\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]} \quad \textbf{Equation 23.}$$

   for $x \geq 0$, where

$$I(\mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \int_0^\infty e^{\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]}. \quad \textbf{Equation 24.}$$

   The cumulative distribution function of the single truncated normal distribution is given by

$$Q(x; \mu, \sigma) = \int_0^x \frac{1}{I(\mu, \sigma)\sigma\sqrt{2\pi}} e^{\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]}\, dx. \quad \textbf{Equation 25.}$$

   The mixture of two-component truncated normal distributions has the following probability density function:

$$r(x; \mu_1, \sigma_1, \mu_2, \sigma_2, w) = wq(x; \mu_1, \sigma_1) + (1-w)q(x; \mu_2, \sigma_2). \quad \textbf{Equation 26.}$$

   Its cumulative distribution function is given by

$$R(x; \mu_1, \sigma_1, \mu_2, \sigma_2, w) = wQ(x; \mu_1, \sigma_1) + (1-w)Q(x; \mu_2, \sigma_2). \quad \textbf{Equation 27.}$$

5. Mixture normal and Weibull distribution

   The mixture normal and Weibull distribution combines a single-truncated normal and a Weibull distribution, which has the following probability density function

$$s(x; \mu, \sigma, k, c) = wq(x; \mu, \sigma) + (1-w)f(x; k, c). \quad \textbf{Equation 28.}$$

   The cumulative distribution is given by

$$S(x; \mu, \sigma, k, c) = wQ(x; \mu, \sigma) + (1-w)F(x; k, c). \quad \textbf{Equation 29.}$$

6. Mixture Weibull and GEV distribution

   The probability density function of a mixture Weibull and GEV distribution is given by the following:

$$t(x;k,c,\zeta,\delta,l) = wf(x;k,c) + (1-w)e(x;\zeta,\delta,l). \quad \textbf{Equation 30.}$$

   The mixture Weibull and GEV cumulative distribution is given by

$$T(x;k,c,\zeta,\delta,l) = wF(x;k,c) + (1-w)E(x;\zeta,\delta,l). \quad \textbf{Equation 31.}$$

7. Mixture Weibull and log-normal distribution

   A random variable X has mixture Weibull and log-normal distribution when it has the following probability density function:

$$u(x;k,c,\lambda,\theta) = wf(x;k,c) + (1-w)l(x;\lambda,\theta). \quad \textbf{Equation 32.}$$

   Its cumulative distribution function is given by

$$U(x;k,c,\lambda,\theta) = wF(x;k,c) + (1-w)L(x;\lambda,\theta). \quad \textbf{Equation 33.}$$

8. Mixture GEV and log-normal distribution The probability density function of the mixture GEV and log-normal distribution is given by

$$v(x;\zeta,\delta,l,\lambda,\theta) = we(x;\zeta,\delta,l) + (1-w)l(x;\lambda,\theta). \quad \textbf{Equation 34.}$$

   The mixture GEV and log-normal distribution has the following cumulative distribution function:

$$V(x;\zeta,\delta,l,\lambda,\theta) = wE(x;\zeta,\delta,l) + (1-w)L(x;\lambda,\theta). \quad \textbf{Equation 35.}$$

Estimation Method

We estimate the parameters of all models using the maximum likelihood estimation method. For all models considered, in the first step, the log likelihood function is written using the probability density functions provided. As an example, the log likelihood function for the mixture of gamma and Weibull distribution can be written as:

$$LL = \sum_{i=1}^{n} \ln\{wg(x,\alpha,\beta) + (1-w)f(x;k,c)\}. \quad \textbf{Equation 36.}$$

In the second step, the GenSA function from R package 'GenSA' is used for optimization.

Selection Criteria

In order to assess the descriptive ability of multiple distributions, we use the following selection criteria: the Akaike Information Criterion, the Bayesian Information Criterion, and the Kolmogorov-Smirnov test.

Akaike Information Criterion (AIC)

AIC is an index to measure the fit of the model and compare the proposed models to other competitive models. It is defined as

$$\mathrm{AIC} = 2k - 2\ln(L) \quad \textbf{Equation 37.}$$

where k is the number of estimable parameters, and $\ln(L)$ is the log-likelihood at its maximum point of the model estimated. While comparing different models, the one with the smallest AIC value is considered the best.

Bayesian Information Criterion (BIC)

BIC is another criterion for model comparison, which is defined as

$$\mathrm{BIC} = \ln(n)k - 2\ln(L) \quad \textbf{Equation 38.}$$

where n is the sample size, k is the number of estimable parameters, and L is the maximum value of the likelihood function. Similar to AIC, the models with smaller BIC value are better than others.

Kolmogorov-Smirnov test (K-S test)

The K-S test is used to check the goodness of fit of a given set of data to a theoretical distribution $F(x)$. Suppose $X_1, X_2, \ldots, X_n$ is a random sample. The null hypothesis, H_o, gives a theoretical distribution function, F_o. The K-S test compares $F_o(x)$, to $S(x)$, the empirical distribution function, where

$$S(x) = \frac{\text{Number of observations with } x_i < x}{n}. \qquad \textbf{Equation } 39.$$

The test statistic is the the maximum (denoted by "sup" for supremum) vertical distance between the two functions and is defined as

$$D = |F_o(x) - S(x)|. \qquad \textbf{Equation } 40.$$

The value of D is compared with a critical value and H_o is rejected for large value of D. For further detail we refer to Conover.[16]

RESULTS AND DISCUSSIONS

In this section, the various models discussed above are applied to two BMI data sets. The first BMI data set was retrieved from the University of Wisconsin–Eau Claire (UWEC)'s Student Health Services. The second BMI data set was collected from National Health and Nutrition Survey (NHANES) and is available in the R package mixsmsn. The model selection criteria, AIC and BIC, and the K-S goodness of fit test p-values are calculated to assess the suitability of the fitted distributions. All computations are conducted using the statistical software R.

UWEC BMI Data

The UWEC BMI data was retrieved from the National College Health Assessment conducted by the UWEC Student Health Service. The analyzed data came from a sample of 630 students attending UWEC during the 2014–2015 academic year. The BMIs of the students who have visited the Student Health Service Center were used for the analysis. **Table 1** shows descriptive statistics for this data set where g_1 and g_2 are the skewness and kurtosis, respectively.

n	$\bar{x}$	s	g_1	g_2
630	24.65	5.07	1.99	8.76

Table 1. Summary Statistics of the UWEC BMI data.

The data clearly shows a highly skewed pattern. Therefore, some recently developed skewed distributions like skew-t and skew-normal distributions may provide competitive fits for this unimodal, skewed data.

We fit all the unimodal and bimodal distributions considered above to this data. The maximum likelihood estimates of the parameters are provided in **Table 3**, found in the Appendix. The AIC and BIC values are presented in **Table 5**. Among the unimodal skew distributions, GEV, skew-t and skew-normal distributions are the best models. When flexible skew distributions are considered, the alpha skew-Laplace distribution is shown to best model the data as it has the smallest AIC and BIC values* (3606.54 and 6315.84, respectively). Weibull, log-normal and gamma distributions are not able to describe BMI characteristics well. This is shown by the small p-value of the K-S test, provided in **Table 5**. Using the estimates of the parameter values from **Table 3**, we plotted the expected densities for all models including the observed data in **Figure 6**.

*See Appendix **Table 5** for complete data.

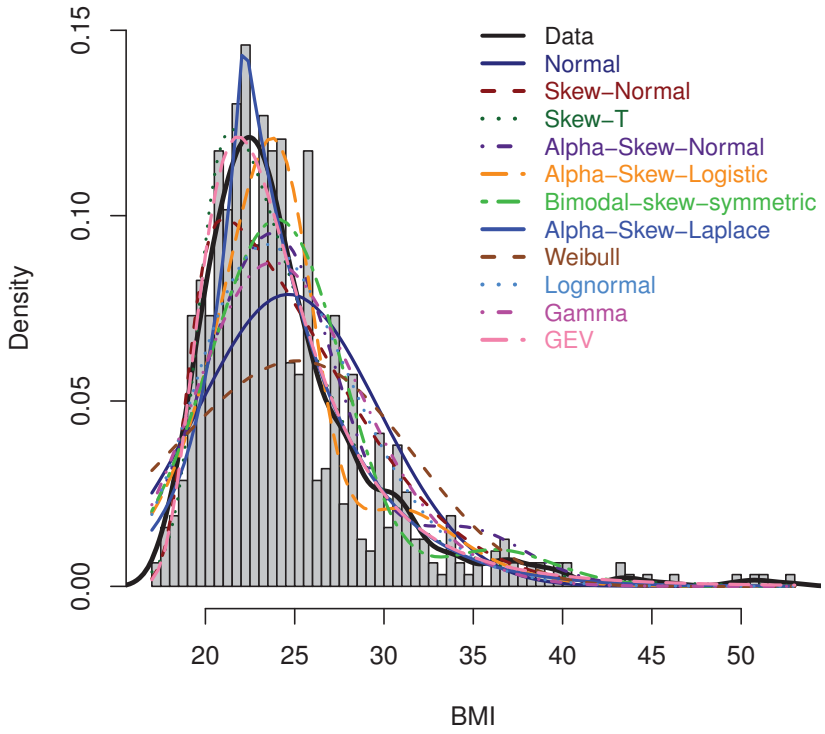

Figure 6. Observed and Expected Density plot of BMI data obtained from 630 UWEC students.

From the observed and expected density plot, it is also confirmed that skew-normal, skew-t, and alpha skew-Laplace models are the best among the skewed and flexible models. Considering the noise at the tail of the observed density, one may argue that flexible distributions, such as bimodal-skew-symmetric distributions and alpha-skew-logistic distributions, also give better fit because they take into account some of that noise. Using the same reasoning, we decided to fit mixture models to this data to see if they provide better fit than the regular models. Using the estimates of the parameter values from **Table 4**, we plotted the expected densities for all mixture models including the observed data in **Figure 7**.

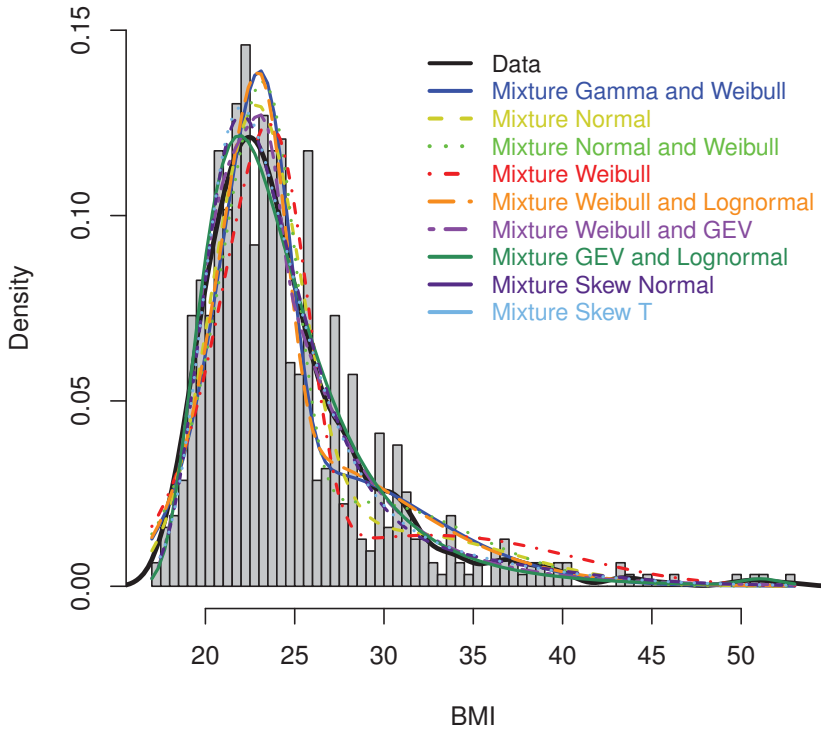

Figure 7. Observed and Expected Density plot of UWEC BMI data using mixture distributions.

Even though univariate distributions and flexible skew distributions show good fits for the UWEC BMI data, the mixture models fit the data better due to their smaller model selection criteria values and larger p-values for the K-S test. According to **Table 6** in the Appendix, mixture of GEV and log-normal provides the closest fits for the observed data followed by mixture of GEV and Weibull, and mixture of scale normal of skew-t .

NHANES BMI Data

We considered the body mass index for men aged between 18 to 80 years. The data set comes from the National Health and Nutrition Examination Survey (NHANES), made by the National Center for Health Statistics (NCHS) of the Center for Disease Control (CDC) in the USA. The data are taken from the NHANES 1999–2000 and NHANES 2001–2002 cycles. Originally, the set had 4579 participants with BMI records. However, to explore the pattern of bi-modality, we consider only those participants who have their weights within [39.50 kg, 70.00 kg] and [95.01 kg, 196.80 kg]. This data set has been used by many authors, for example Lin et al.,[8] and is available in the R package mixsmsn. BMI is defined as the ratio of body weight in kilograms and body height in meters squared. **Table 2** shows descriptive statistics for this data set.

n	$\bar{x}$	s	g_1	g_2
2107	28.19	7.50	0.71	3.30

Table 2. Summary Statistics of the body mass index of 2107 American men.

The mean and standard deviation of the NHANES BMI data are 28.19 and 7.50, respectively. The skewness of the data is 0.71, and the kurtosis is 3.30. The data clearly shows two peaks with some skew pattern. Therefore, some recently developed

flexible skewed distributions such as alpha skew-normal distribution, alpha-skew-logistic distribution, alpha-skew-Laplace distribution may provide more competitive fit for this bimodal skewed data.

We fit all of the skewed and flexible skewed distributions considered above to the NHANES BMI data. Since skew-t and skew-normal distributions most accurately model the UWEC BMI data, we also consider the mixture of these distributions to describe the NHANES BMI data. The maximum likelihood estimates of the parameters are provided in **Table 3**, found in the Appendix. The AIC and BIC values are presented in the **Table 5**. According to AIC and BIC values, among the unimodal skewed distributions, skew-t and skew-normal distributions are the best models while other traditional skewed distributions such as gamma, log-normal, and Weibull has relatively high AIC and BIC values. Using the estimates of the parameter values from **Table 3**, we plotted the expected densities for all models including the observed data in **Figure 8**. Although skew-t and skew-normal are a better fit for the NHANES data than the other univariate distributions, they do not account for the second peak of the distribution as observed from **Figure 8**.

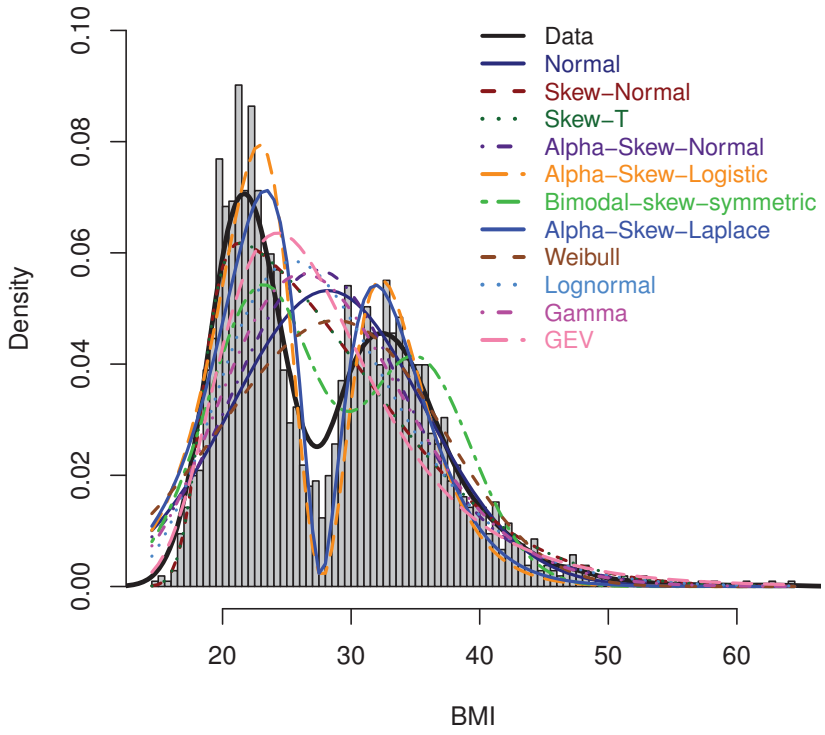

Figure 8. Observed and expected densities of NHANES BMI Data with skewed models.

Among the flexible skewed distributions, we find evidence that alpha-skew-Laplace best describes the NHANES data characteristics—due to its small model selection criteria values (AIC and BIC values)—followed by alpha skew-logistic distribution. These distributions not only have small AIC and BIC values but they also take into account the second peak of the distribution. Since the data are bimodal, we believe that mixture distributions may be suitable to model the data as well. The maximum likelihood estimates for the parameters of the mixture models are presented in **Table 4** in the appendix. The model selection criteria, AIC and BIC values, are given in **Table 6**.

According to the AIC and BIC, mixture of GEV and Log-normal model turns out to be the best fitting model (AIC = 13720.74, BIC = 13724.05). Mixture of skew-normal and mixture of skew-t also describe the data well with small AIC

values of 13764.42 and 13737.61, respectively, and BIC values of 13803.99 and 13777.19, respectively. If we consider large *p*-values from the K-S test with the AIC and BIC, then the mixture of gamma and Weibull also provides a good fit.

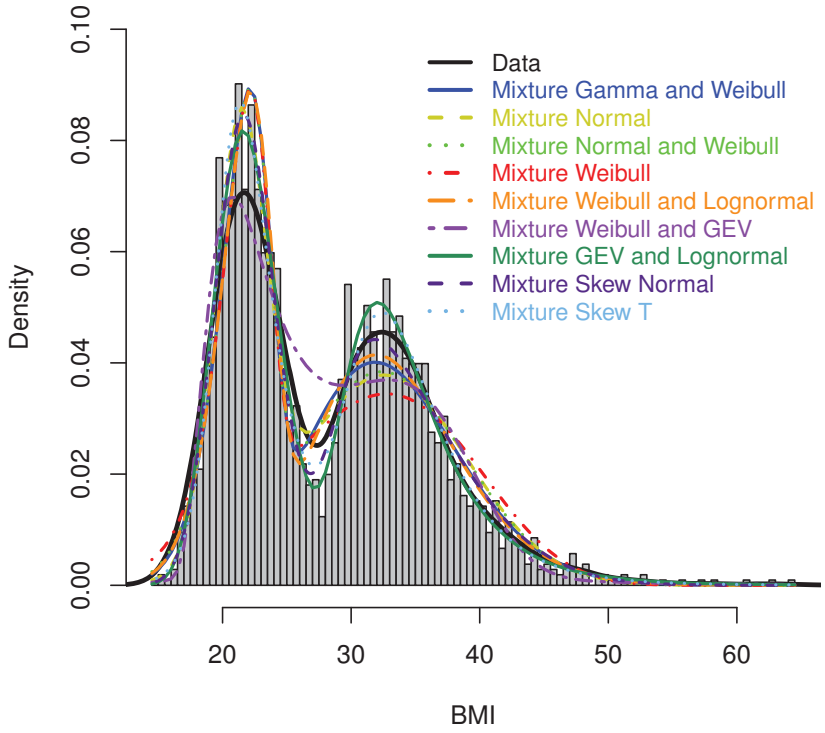

Figure 9. Observed and expected density plot of NHANES BMI Data using mixture distributions

Figure 9 shows the observed and expected density plot of all the mixture models. From the density plot we clearly observed that the scale mixture of skew-normal and skew-t distributions fit the data reasonably well along with the mixture of GEV and log-normal distributions. One of the main advantages of using flexible skewed distributions over mixture distributions is that they are easier to interpret because of the smaller number of parameters. For the mixture models, as the number of components increase, the parameters become very difficult to interpret.

CONCLUSIONS

In this paper, we demonstrated a framework for testing features, such as bimodality, asymmetry, and robustness, of BMI data. We conducted a comparison of skewed, flexible and mixture distribution models using two sets of BMI data. The comparison of the distributions is made using AIC, BIC and the K-S test. We not only looked at conventional unimodal distributions, but also included recently constructed flexible skewed distributions and the models of their mixture. Our findings showed that among the univariate distributions, skew-t and skew-normal proved to provide the best fit for both of the data sets. However, these unimodal skewed distributions fail to account for the second mode of the NHANES BMI data. Among the flexible skewed distributions, alpha skew-Laplace distribution best describes both data sets. Though univariate distributions and flexible skewed distributions may have shown a fairly accurate description of the BMI data, there is evidence that the mixture distributions have considerably good fit as well. For the UWEC BMI data, mixture of GEV and log-normal and the scale mixture of skew-normal and skew-t provided the best fit out of all considered distributions. For the NHANES BMI data, mixture of GEV and log-normal, and the scale mixture of skew-t are the most suitable distributions to represent the data. There are no significant differences between the performance of mixture of GEV and log-normal,

mixture of skew-normal, and mixture of skew-t—though mixture of GEV and log-normal performed slightly better. As a standard health indicator, BMI data still needs extended studies beyond this framework. Given other data sets, we can further analyze the covariance relationship between BMI and other variables such as age and gender using a regression model with error distributions discussed in this paper which covers both skewness and heavy tailed properties of some real data. A more accurate understanding of BMI distribution can assist further research involving topics like obesity and its links to health outcomes.

ACKNOWLEDGMENTS

We want to thank the Office of Research and Sponsored Programs (ORSP) and the Mathematics Department at the University of Wisconsin–Eau Claire for supporting our work, and Student Health Services at UWEC for providing us the data. We also want to thank two anonymous referees for their valuable comments which improved the standard of the paper.

REFERENCES

1. Ma, Y., Genton, M. G., Tsiatis, A. A. (2005). Locally efficient semiparametric estimators for generalized skew-elliptical distributions. *Journal of the American Statistical Association*, **100**(471), 980–989.
2. Canale, A. (2011). Statistical aspects of the scalar extended skew-normal distribution. *International Journal of Statistics*, **69**(3), 279, 295.
3. Olivares-Pacheco, J. F., Salas, E., Gómez, H. W., Bolfarine, H. (2012). An asymmetric extension for the family of elliptical slash distributions. *Communications in Statistics - Theory and Methods*, **41**, 1000–1012.
4. Tan, F., Tang, Y., Peng, H. (2015). The multivariate slash and skew-slash student t distributions. *Journal of Statistical Dirstribution and Applications*, **2**(1), 1–22.
5. Arslan, O. (2015). Variance-mean mixture of the multivariate skew-normal distribution. *Statistical Papers*, **56**(2), 353–378.
6. Manyema, M., Veerman, L. J., Chola, L., Tugendhaft, A., Sartorius, B., Labadarios, D., Hofman, K. J. (2014). The potential impact of a 20% tax on sugar-sweetened beverages on obesity in South African adults: A mathematical model. *PLoS ONE*, **9**(8).
7. Miljkovic, T., Shaik, S., Miljkovic, D. (2016). Redefining standards for body mass index of the US population based on BRFSS data using mixtures. *Journal of Applied Statistics*, 1–15.
8. Lin, T., Lee, J., Hsieh, W. (2007). Robust mixture modeling using the skew-t distribution. *Statistics and Computing*, **17**(2), 81–92.
9. Elal-Olivero, D. (2010). Alpha-skew-normal distribution. *Proyecciones*, **29**, 224–240.
10. Harandi, S. S., Alamatsaz, M. H. (2012). Alpha-Skew-Laplace distribution. *Statistics and Probability Letters*, **83**, 774–782.
11. Hazarika, P. J., Chakraborty, S. (2014). Alpha-Skew-Logistic Distribution. *IOSR Journal of Mathematics*, **10**, 36–46.
12. Hassan, M.Y., El-Bassiouni, M.Y. (2013). Bimodal Skew-Symmetric Normal Distribution. *Communications in Statistics–Theory and Methods*, **45**(5), 1527–1541.
13. Kollu, R., Rayapudi, S. R., Narasimham, S. V. L., Pakkurthi, K.M. (2012). Mixture probability distribution functions to model wind speed distributions. *International Journal of Energy and Environmental Engineering*, **3**(27), 1–10.
14. Aziz, M. (2011). Study of Unified Multivariate skew-normal Distribution with Applications in Finance and Actuarial Science. Unpublished PhD dissertation, Bowling Green State University.
15. Johnson, N. L., Kotz, S., & Balakrishnan, N. (1994). *Continuous univariate distributions*. New York: Wiley.
16. Conover, W. J.,(1999). *Practical Nonparametric Statistics, 3rd edition*, John Wiley & Sons, Inc. New York.
17. Azzalini, A. (1985). A class of distributions which includes the normal ones. *Scandinavian Journal of Statistics*, **12**, 171–178.
18. Wang, J., Boyer, J., Genton M. G. (2004). A skew-symmetric representation of multivariate distributions. *Statistica Sinica.*, **14**, 1259–1270.
19. Ma, Y., Genton, M. G. (2004). Flexible class of skew-symmetric distributions, *Scandinavian Journal of Statistics*, **31**, 459–468.

20. Basso, R. M., Cabral, C. R. B., Lachos, V. H., Ghosh, P. (2010). Robust mixture modeling based on scale mixtures of skew-normal distributions. *Computational Statistics & Data Analysis*, **54**(12), 2926–2941.

ABOUT THE STUDENT AUTHORS

Thao Tran is an undergraduate student at the University of Wisconsin–Eau Claire with a double major in Comprehensive Finance and Mathematics with a Statistics emphasis. She is also a member of Kappa Mu Epsilon (a mathematics honor society) and the University Honors Program. With her high impact practices, she plans to expand her knowledge and conduct further research in Statistics at the graduate level.

Cara Wiskow is a 3rd year undergraduate student at the University of Wisconsin–Eau Claire and is pursuing a degree in Mathematics with a Statistics emphasis. She is a member of Kappa Mu Epsilon and the University Honors Program. She is also the student representative on the university's Institutional Review Board. After completing her undergraduate coursework, Cara plans on attending graduate school for an MPH in Biostatistics.

PRESS SUMMARY

The normal distribution comes as a first choice while fitting real data, but it may not be suitable to use if the assumed distribution deviates from normality. Flexible skewed distributions have recently drawn considerable attention as alternative models because they are not only capable of including skewness but also flexible enough to take into account multimodality. Using two BMI data sets, we used flexible skewed and mixture distributions to search for the best models. Our results indicate that the skew-t and alpha-skew-Laplace distributions are reasonably competitive when describing unimodal BMI data whereas mixture of normal and finite mixture of scale mixture of skew-normal and skew-t distributions are better alternatives to both unimodal and bimodal conventional distributions. We believe the models discussed in ours study will offer a framework for testing features such as bimodality, asymmetry, and robustness of the BMI data, thus providing a more detailed and accurate understanding of the distribution of BMI.

APPENDIX

Model		NHANES BMI Data	UWEC BMI
Normal	μ	28.188	24.649
	σ	7.499	5.065
Skew-normal	μ	28.151	18.937
	σ	7.546	7.634
	α	0.951	8.048
Skew-t	μ	18.314	19.575
	σ	12.397	5.241
	α	9.454	4.446
	ν	16896.843	4.198
Weibull	σ	31.067	4.296
	α	3.897	26.751
GEV	μ	24.556	0.1631
	σ	0.0453	3.070
	α	5.791	22.294
Log-normal	μ	3.305	3.187
	σ	0.259	0.182
Gamma	α	14.915	28.314
	β	0.529	1.149
Alpha-skew-normal	μ	32.314	27.987
	σ	7.123	4.335
	α	0.754	1.357
BSSN	μ	28.891	28.500
	ψ	0.019	0.024
	β	29.491	32.434
	δ	18.191	5.864
Alpha-skew-logistic	μ	27.310	26.314
	α	4.571	0.857
	β	1.968	1.737
Alpha-skew-Laplace	μ	27.332	22.140
	σ	2.146	2.960
	α	7.307	-0.355

Table 3. Estimated parameter values for skew and flexible models.

Model		NHANES BMI Data	UWEC BMI
MN	μ_1	32.548	30.581
	σ_1	6.418	6.514
	μ_2	21.413	22.809
	σ_2	2.018	2.480
	ω	0.608	0.237
MSN	μ_1	28.457	19.871
	σ_1	62.240	19.215
	α_1	3.667	2.869
	μ_2	25.529	20.221
	σ_2	7.997	94.663
	α_2	0.991	3.661
	ω	0.517	0.848
MST	μ_1	19.877	19.928
	σ_1	9.198	18.240
	α_1	1.444	2.708
	μ_2	29.742	26.078
	σ_2	33.249	83.404
	α_2	2.246	3.296
	ν	6.021	25.501
	ω	0.517	0.859
MW	σ_1	34.427	32.355
	α_1	4.760	4.495
	σ_2	22.203	23.746
	α_2	13.121	10.417
	ω	0.661	0.279
MNW	μ_1	32.377	28.865
	σ_1	6.456	6.158
	σ_2	22.166	23.316
	α	12.487	11.823
	ω	0.622	0.351
MGW	α_1	30.479	23.704
	β	0.925	0.850
	σ	22.169	23.0980
	α_2	12.157	12.640
	ω	0.592	0.429
MWLN	σ_1	22.258	13.138
	α	11.924	22.962
	μ	0.169	0.201
	σ_2	3.493	3.288
	ω	0.430	0.516

Model		NHANES BMI Data	UWEC BMI Data
MWGEV	σ_1	35.131	23.449
	α_1	6.662	30.108
	μ	21.442	0.172
	σ_2	3.253	3.156
	α_2	0.285	22.296
	ω	0.438	0.056
MGEVLN	μ_1	32.400	22.295
	σ_1	3.520	3.038
	α	0.088	0.139
	μ_2	3.087	3.934
	σ_2	0.122	0.021
	ω	0.483	0.995

Table 4. Estimated parameter values for mixture models

Model	NHANES BMI Data				UWEC BMI Data			
	LogL	AIC	BIC	K-S Test	LogL	AIC	BIC	K-S Test
Normal	-7234.19	14472.38	14483.69	2.20E-16	-1915.99	3835.98	3844.87	2.47E-12
Skew-normal	-6995.39	13996.78	14013.74	2.20E-16	-1790.70	3587.41	3594.30	1.27E-6
Skew-t	-6995.40	13982.79	13975.49	1.64E-10	-1763.31	3534.62	3539.51	0.4050
Weibull	-7280.88	14565.76	14577.07	2.20E-16	-1988.89	3981.80	3990.69	2.20E-16
GEV	-7092.59	14191.18	14200.48	2.20E-16	-1759.766	3525.53	3532.423	2.20E-16
Log-normal	-7104.323	14212.65	14223.95	2.20E-16	-1827.00	3657.99	3666.88	2.20E-16
Gamma	-7130.11	14264.22	14275.52	2.20E-16	-1852.21	3708.42	3717.31	2.20E-16
Alpha-skew-normal	-7202.510	14399.02	14427.98	2.20E-16	-1840.63	3687.27	3694.16	2.20E-16
BSSN	-7141.82	14291.64	14298.94	NA †	-1833.82	3675.63	3680.53	NA
Alpha-skew-logistic	-7128.22	14262.44	14271.75	NA	-1825.67	3657.34	3664.23	NA
Alpha skew-Laplace	-7091.47	14188.93	14198.24	2.20E-16	-1800.27	3606.54	3615.84	2.20E-16

Table 5. Model fitting results: skewed and flexible models.

Model	NHANES BMI Data				UWEC BMI Data			
	LogL	AIC	BIC	K-S Test	LogL	AIC	BIC	K-S Test
MN	-6911.67	13833.35	13803.66	2.20E-16	-1779.75	3549.50	3546.61	2.20E-16
MSN	-6875.21	13764.42	13803.99	NA	-1758.29	3534.25	3565.37	NA
MST	-6859.81	13737.61	13777.19	NA	-1758.29	3534.25	3565.37	NA
MW	-6984.54	13979.08	13984.38	3.29E-4	-1814.92	3639.84	3642.7	2.20E-16
MNW	-6926.70	13863.39	13868.70	0.0262	-1801.98	3613.95	3616.84	0.2480
MGW	-6903.83	13817.66	13822.96	0.1390	-1787.58	3585.16	3588.05	0.3490
MWLN	-6897.41	13804.82	13810.12	0.2682	-1781.41	3574.82	3575.71	0.2873
MWGEV	-6938.18	13888.36	13891.67	0.00122	-1758.21	3528.43	3529.32	2.20E-16
MGEVLN	-6854.37	13720.74	13724.05	2.20E-16	-1757.57	3527.13	3528.03	2.20E-16

Table 6. Model fitting results: mixture models.

†Not available