Maximum Likelihood Estimation of Dietary Intake Distributions

Jeffrey D. Helterbrand

Working Paper 92-WP 98
August 1992

Center for Agricultural and Rural Development
Iowa State University
Ames, Iowa 50011

Jeffrey D. Helterbrand is a graduate assistant, Department of Statistics, Iowa State University, Ames, Iowa.

This paper was prepared for the Human Nutrition Information Service of USDA under Research Support Agreement 58-3198-9-032.

ABSTRACT

This paper applies maximum likelihood estimation techniques to determine suitable models for dietary intake distributions. Hypothesis test results indicate that while the gamma and Weibull models appear suitable for describing the intake distributions of some dietary components, a more flexible family of distributions is required in order to appropriately encompass all dietary component distributions.

Six nutrients are considered in the analysis including calcium, energy, iron, protein, vitamin A, and vitamin C. Based on chi-square goodness-of-fit tests, we conclude that the three parameter, generalized gamma family of distributions accurately describes the distributions of all six dietary components.

The additional flexibility of this family results in large standard errors for the parameter estimates. However, the standard errors of the estimated percentage of the population below a specified level of nutrient intake appear precise and allow for substantive conclusions regarding nutritional inadequacy to be made.

Introduction

An estimate of the distribution of intakes for a given dietary component can be used to obtain estimates of the prevalence of inadequacy in the nutritional status of individuals. Dietary inadequacies may result from either excessive or deficient intakes of a dietary component over a long period of time. Dietary intake distributions are of interest to nutritionists and policy makers for monitoring dietary and nutritional status, for evaluating food policies, and in nutrition and dietary education programs.

Estimating dietary intake distributions typically involves estimating distributions of usual intake. The usual daily intake of a dietary component for an individual is a measure of the individual's typical daily consumption rate during a time period appropriate for the particular dietary component. Usual intake can be viewed as a long-run average of daily intakes for the individual.

There is clear evidence that many intake distributions are not symmetric. Hence, distributions such as the gamma and Weibull family are potential models for usual intake distribution fitting. In preliminary investigations, Nusser et al. (1988) indicated that assuming that usual intakes follow a gamma distribution was appropriate for some, but not all, dietary components. A similar result was obtained using the Weibull family. Thus, Nusser et al. suggested the generalized gamma distribution as a potentially all-encompassing family of distributions for dietary component distributions.

Properties of the Generalized Gamma Distribution

The generalized gamma family is a three parameter distribution with density

$$f(y) = [\theta\Gamma(\beta)]^{-1}\lambda(\theta^{-1}y)^{\lambda\beta-1}\exp\{-(\theta^{-1}y)^{\lambda}\} , \qquad (1)$$

$$y > 0 , \ \theta > 0 , \ \beta > 0 , \ \lambda > 0 ,$$

where $\Gamma(\beta) = \int_0^\infty t^{\beta-1}e^{-t}dt$ and θ , β , and λ are parameters to be estimated. The generalized gamma family has many familiar distributional families as special cases. Examples include the exponential $(\beta = 1, \lambda = 1)$, gamma $(\lambda = 1)$, Weibull $(\beta = 1)$ and chi-square $(\theta = 2, \beta = n/2, \lambda = 1)$ distributions. In addition, the lognormal family is a limiting special case $(\beta \to \infty)$.

The cumulative distribution function for the generalized gamma distribution is

$$F(y; \theta, \beta, \lambda) = \Gamma_Z(\beta)/\Gamma(\beta) , \qquad (2)$$

where $Z = (y/\theta)^{\lambda}$ and $\Gamma_Z(\beta) = \int_0^Z t^{\beta-1}e^{-t}dt$. The r-th moment of Y $(r=1, 2, 3, \ldots)$ can be written

$$E(Y^r) = \theta^r\Gamma^{-1}(\beta)\ \Gamma(\beta + r/\lambda) . \qquad (3)$$

If Y is a generalized gamma variate, then Y^{λ} is distributed as a gamma with parameters θ^{λ} and β . This property of the generalized gamma distribution is used in obtaining starting estimates for the iterative procedures necessary in estimating generalized gamma parameters.

The generalized gamma distribution is a more general model than the Weibull, exponential, and gamma. However, Hager and Bain (1970) concluded from their study that the Weibull model was about as flexible as the generalized gamma distribution for sample sizes up to 200. Thus, given the complexity of the generalized gamma distribution, and some of the estimation difficulties encountered, Hager and Bain suggested that the Weibull assumption was preferable to a generalized gamma distribution for sample sizes up to 200.

Maximum Likelihood Estimation and the Generalized Gamma Distribution

Stacy and Mihram (1965), Hager and Bain (1970), Parr and Webster (1965), and Prentice (1974) have examined maximum likelihood techniques to estimate the parameters of the generalized gamma distribution. From the density function, the maximum likelihood equations for n independent observations can be obtained as follows. The likelihood function is

$$f(y_1, y_2, \ldots, y_n; \theta, \beta, \lambda) = \prod_{i=1}^{n} [\theta\Gamma(\beta)]^{-1}\lambda(\theta^{-1}y_i)^{\lambda\beta-1}\exp\{-(\theta^{-1}y_i)^{\lambda}\} . \tag{4}$$

Let $L(n) = \ell n\, f(y_1, y_2, \ldots, y_n; \theta, \beta, \lambda)$ be the log-likelihood function. It follows that

$$L(n) = n\,\ell n(\lambda) - n\lambda\beta\,\ell n(\theta) - n\,\ell n[\Gamma(\beta)] + (\lambda\beta - 1)\sum_{i=1}^{n}\ell n(y_i) - \frac{1}{\theta^{\lambda}}\sum_{i=1}^{n} y_i^{\lambda} . \tag{5}$$

The maximum likelihood equations are obtained by taking the first derivative of $L(n)$ with respect to θ, β and λ, respectively, and setting the resulting derivatives equal to zero. The equations are

$$-n\beta + \sum_{i=1}^{n} (y_i/\theta)^{\lambda} = 0 , \tag{6}$$

$$-n\psi(\beta) + \lambda \sum_{i=1}^{n} \ell n(y_i/\theta) = 0 , \tag{7}$$

$$n/\lambda + \beta \sum_{i=1}^{n} \ell n(y_i/\theta) - \sum_{i=1}^{n} (y_i/\theta)^{\lambda} \ell n(y_i/\theta) = 0 , \tag{8}$$

where $\psi(\beta) = d[\ell n\Gamma(\beta)]/d\beta$. Equations (6), (7), and (8) must be solved for θ, β, and λ simultaneously. Since a closed form solution for θ, β, and λ is not known, an iterative technique is required to compute the estimators $\hat{\theta}$, $\hat{\beta}$, and $\hat{\lambda}$.

Using maximum likelihood techniques to estimate the generalized gamma parameters requires good starting values for the parameters in order to obtain the appropriate estimates from the iterative procedure. Adequate computer resources are also needed to calculate functions such as $\Gamma(\beta)$, $\psi(\beta)$ and $\psi'(\beta)$. Harter (1965) found the number of iterations required for convergence tended to be large when estimating β and λ simultaneously. This is apparently due to the high negative correlation between the estimates of the two parameters. In addition, the approximate normal distribution for $\hat{\beta}$ predicted by maximum likelihood theory was not observed even for samples of size 400 [Prentice (1974)].

To apply maximum likelihood, it is necessary to develop a reliable iterative technique which will converge to correct solutions, while keeping the estimates in the parameter space $(\theta > 0, \beta > 0, \lambda > 0)$.

Hager and Bain (1970) indicated that the Newton-Raphson method did not appear to work well when solving for generalized gamma parameters. Harter (1965) advised using a hybrid of two different iterative techniques to find parameter estimates which converge.

An Algorithm for Maximum Likelihood Estimation for the Generalized Gamma Distribution

An algorithm for estimating generalized gamma parameters for the usual intake data, requires two stages: 1) an algorithm providing good starting values and 2) an iterative algorithm to compute accurate solutions from the given starting values.

To obtain estimates of the generalized gamma parameters, equations for the first four central moments of the generalized gamma distribution are needed. They are

$$\mu_1 = E\{y\} = \theta\Gamma^{-1}(\beta)\Gamma(\beta + \lambda^{-1}) \ , \tag{9}$$

$$\mu_2 = E\{(y - \mu)^2\} = \theta^2\{\Gamma^{-1}(\beta)\Gamma(\beta + 2\lambda^{-1}) - [\Gamma^{-2}(\beta)\Gamma^2(\beta + \lambda^{-1})]\} \ , \tag{10}$$

$$\mu_3 = E\{(y - \mu)^3\} = \theta^3[\Gamma^{-1}(\beta)\Gamma(\beta + 3\lambda^{-1}) - 3\Gamma^{-2}(\beta)\Gamma(\beta + \lambda^{-1})\Gamma(\beta + 2\lambda^{-1})$$

$$+ 2\Gamma^{-3}(\beta)\Gamma^3(\beta + \lambda^{-1})] \ , \tag{11}$$

$$\mu_4 = E\{(y - \mu)^4\} = \theta^4[\Gamma^{-1}(\beta)\Gamma(\beta + 4\lambda^{-1})$$

$$- 4\Gamma^{-2}(\beta)\Gamma(\beta + \lambda^{-1})\Gamma(\beta + 3\lambda^{-1})$$

$$+ 6\Gamma^{-3}(\beta)\Gamma^2(\beta + \lambda^{-1})\Gamma(\beta + 2\lambda^{-1})$$

$$- 3\Gamma^{-4}(\beta)\Gamma^4(\beta + \lambda^{-1})] \quad . \tag{12}$$

The starting value routine begins by using a grid search scheme. Fifteen λ values are chosen from a range of values believed to contain the parameter. For a given power, λ, a starting value for the shape parameter, β_λ, is obtained by solving the second moment equation (10) evaluated at λ. An iterative procedure for obtaining a solution for β_λ from (10) is the DBCPOL routine in IMSL, which also requires a starting value for β_λ. Since $y \sim GG(\theta, \beta, \lambda)$ implies $y^\lambda \sim \text{Gamma}(\theta^\lambda, \beta)$, an initial value for β_λ can be derived as follows. Using the second order Taylor series expansion, the mean and variance of y^λ can be approximated by

$$E(y^\lambda) \doteq \mu_1^\lambda + a_2\mu_2 = \mu_{1\lambda} \tag{13}$$

and

$$\text{Var}(y^\lambda) \doteq a_1^2\mu_2 + 2a_1a_2\mu_3 + a_2^2(\mu_4 - \mu_2^2) = \mu_{2\lambda} \quad , \tag{14}$$

where

$$a_1 = \lambda\mu_1^{\lambda-1} \quad ,$$

$$a_2 = 2^{-1}\lambda(\lambda - 1)\mu_1^{\lambda-2} ,$$

and μ_1 , μ_2 , μ_3 and μ_4 are the moments of y . Since y^λ follows a gamma distribution,

$$\beta = [\text{var}(y^\lambda)]^{-1}[E(y^\lambda)]^2 . \tag{15}$$

Hence an initial value for β_λ is

$$\hat{\beta}_{\lambda 0} = \hat{\mu}_{2\lambda}^{-1}\hat{\mu}_{1\lambda}^2 , \tag{16}$$

where $\hat{\mu}_{1\lambda}$ and $\hat{\mu}_{2\lambda}$ are evaluated at the sample moments $\hat{\mu}_1$, $\hat{\mu}_2$, $\hat{\mu}_3$, and $\hat{\mu}_4$. Thus for each λ , $\hat{\beta}_{\lambda 0}$ is used to start the DBCPOL subroutine and $\hat{\beta}_\lambda$ is computed. A starting value, $\hat{\theta}_\lambda$ is also calculated by evaluating the first moment equation (9) at λ and $\hat{\beta}_\lambda$.

The starting value algorithm computes fifteen vectors, $(\lambda, \hat{\beta}_\lambda, \hat{\theta}_\lambda)$, corresponding to the fifteen starting λ values. The algorithm next substitutes each of these vectors into the log-likelihood equation (5) for the generalized gamma distribution, and a value for the log-likelihood is calculated. The three largest log-likelihood values and their corresponding λ values are determined. A quadratic in λ is fit to these three log-likelihood points, and the λ corresponding to the maximum of the quadratic is taken to be the starting value, λ^* , for stage 2. Given λ^* , equation (16) is used to compute $\hat{\beta}_{\lambda^*}$, and the first moment equation (9) is then solved to calculate $\hat{\theta}_{\lambda^*}$. The vector $(\lambda^*, \hat{\beta}_{\lambda^*}, \hat{\theta}_{\lambda^*})$ serves as the starting value for the iterative maximum likelihood estimation program.

To compute the maximum likelihood estimators, the nonlinear system

$$F_1(y;\ \theta,\ \beta,\ \lambda) = -n\beta + \sum_{i=1}^{n} (y_i/\theta)^{\lambda} = 0 \ , \tag{17}$$

$$F_2(y;\ \theta,\ \beta,\ \lambda) = -n\psi(\beta) + \lambda \sum_{i=1}^{n} \ell n(y_i/\theta) = 0 \ , \tag{18}$$

$$F_3(y;\ \theta,\ \beta,\ \lambda) = n/\lambda + \beta \sum_{i=1}^{n} \ell n(y_i/\theta) - \sum_{i=1}^{n} (y_i/\theta)^{\lambda} \ell n(y_i/\theta) = 0 \ , \tag{19}$$

is solved using a modified Newton's method. This system corresponds to (6), (7), and (8). Newton's iterative method requires $J(y;\ \theta,\ \beta,\ \lambda)$, the Jacobian matrix of partial derivatives of F_1, F_2, and F_3 with respect to θ, β, and λ. The elements of this matrix are

$$J_{11}(y;\ \theta,\ \beta,\ \lambda) = -\lambda\theta^{-(\lambda+1)} \sum_{i=1}^{n} y_i^{\lambda} \ , \tag{20}$$

$$J_{12}(y;\ \theta,\ \beta,\ \lambda) = -n \ , \tag{21}$$

$$J_{13}(y;\ \theta,\ \beta,\ \lambda) = (1/\theta^{\lambda}) \left(\sum_{i=1}^{n} y_i^{\lambda} \ell n\ y_i - \ell n\theta \sum_{i=1}^{n} y_i^{\lambda} \right) \ , \tag{22}$$

$$J_{21}(y;\ \theta,\ \beta,\ \lambda) = -n\lambda/\theta \ , \tag{23}$$

$$J_{22}(y;\ \theta,\ \beta,\ \lambda) = -n\psi'(\beta) \ , \tag{24}$$

$$J_{23}(y;\ \theta,\ \beta,\ \lambda) = \sum_{i=1}^{n} \ell n(y_i/\theta) \ , \tag{25}$$

$$J_{31}(y;\ \theta,\ \beta,\ \lambda) = -(n/\theta) + \theta^{-(\lambda+1)} \left\{ \sum_{i=1}^{n} (y_i^{\lambda} \ell n\ y_i - y_i^{\lambda}) + \lambda\ \ell n\ \theta \sum_{i=1}^{n} y_i^{\lambda} \right\} \ ,$$

(26)

$$J_{32}(y;\ \theta,\ \beta,\ \lambda) = \sum_{i=1}^{n} \ell n(y_i/\theta)\ , \tag{27}$$

$$J_{33}(y;\ \theta,\ \beta,\ \lambda) = -(n/\lambda^2) - \theta^{-\lambda}\Big(\sum_{i=1}^{n} y_i^{\lambda}(\ell n\ y_i)^2 - (\ell n\ \theta)^2 \sum_{i=1}^{n} y_i^{\lambda}\Big)\ . \tag{28}$$

The $(n + 1)$-st estimates of the Newton method satisfy the system,

$$\begin{bmatrix} \hat{\theta}_{n+1} \\ \hat{\beta}_{n+1} \\ \hat{\lambda}_{n+1} \end{bmatrix} = \begin{bmatrix} \hat{\theta}_{n} \\ \hat{\beta}_{n} \\ \hat{\lambda}_{n} \end{bmatrix} - J^{-1} \begin{bmatrix} F_1(y;\ \hat{\theta}_n,\ \hat{\beta}_n,\ \hat{\lambda}_n) \\ F_2(y;\ \hat{\theta}_n,\ \hat{\beta}_n,\ \hat{\lambda}_n) \\ F_3(y;\ \hat{\theta}_n,\ \hat{\beta}_n,\ \hat{\lambda}_n) \end{bmatrix},$$

where $(\hat{\theta}_n,\ \hat{\beta}_n,\ \hat{\lambda}_n)$ are the estimates at the n-th iteration. At each iteration n, the log-likelihood is calculated using $(\hat{\theta}_n,\ \hat{\beta}_n,\ \hat{\lambda}_n)$ to assure that the estimates are converging to a solution which maximizes the log-likelihood function. The iteration is assumed to have converged when the maximum of $(|\hat{\theta}_{n+1} - \hat{\theta}_n|,\ |\hat{\beta}_{n+1} - \hat{\beta}_n|,\ |\hat{\lambda}_{n+1} - \hat{\lambda}_n|)$ is less than 0.00001. Once convergence is achieved, the maximum likelihood estimates are denoted by $\hat{\theta}_{MLE}$, $\hat{\beta}_{MLE}$, and $\hat{\lambda}_{MLE}$.

One computational difficulty in this algorithm is the calculation of $\psi(\beta)$ and $\psi'(\beta)$. For real positive β, $\psi(\beta)$ is a concave increasing function which satisfies the following relations (Bernardo, 1976)

$$\psi(1) = -\gamma\ , \tag{29}$$

$$\psi(1 + \beta) = \psi(\beta) + 1/\beta\ , \tag{30}$$

$$\psi(\beta) = \ell n\beta - \frac{1}{2\beta} - \frac{1}{12\beta^2} + \frac{1}{120\beta^4} - \frac{1}{252\beta^6} + 0(\frac{1}{\beta^8}) \quad \text{(for large } \beta \text{ i.e. } \beta \to \infty) \text{ ,} \tag{31}$$

$$\psi(\beta) = -\gamma - \frac{1}{\beta} + 0(\beta) \quad \text{(for small } \beta \text{ i.e. } \beta \to 0) \text{ ,} \tag{32}$$

where γ is Eulers constant. Equation (32) is used to compute $\psi(\beta)$ for $\beta < 1.0 \times 10^{-5}$. The Stirling expansion (31) is used in calculating $\psi(\beta)$ for $\beta \geq 8.5$. The recursive equation (30) is used for $1.0 \times 10^{-5} \leq \beta < 8.5$, and gives values for $\psi(\beta)$ that are accurate to within 1.0×10^{-5} for β in that range. By differentiating these continuous functions, the resulting equations are used to calculate $\psi'(\beta)$ for the three cases.

A second algorithm for fitting the parameters of the generalized gamma distribution can be constructed using a reparameterization of the generalized gamma density based on the logarithms of the response variable. This procedure was suggested by Prentice (1974). The log generalized gamma variate follows a location-scale model. The density function under the Prentice parameterization is

$$f(x;\ \alpha,\ \sigma,\ q) = |q|\{\sigma\Gamma(q^{-2})\}^{-1} \exp(\omega q^{-2} - e^{\omega}) \qquad (q \neq 0)$$

$$= \left(\frac{1}{2\pi\sigma^2}\right)^{1/2} \exp\{-\frac{1}{2\sigma^2}(x - \alpha)^2\} \qquad (q = 0) \text{ ,}$$

where $x = \log y$ and $\omega = \frac{q}{\sigma}(x - \alpha) + \psi(q^{-2})$. Note that, by this reparameterization, the generalized gamma model has been extended such that q can be negative. If $q = \beta^{-1/2}$, $x = \log(y)$ follows a normal model when $q = 0$. The Prentice parameters (parameterization B) can be

expressed in terms of the original parameters (parameterization A) by the nonlinear mapping

$$q = \beta^{-1/2} \quad , \quad \alpha = \log \theta + [\psi(\beta)/\lambda] \quad , \text{ and } \quad \sigma = (1/\lambda\beta^{1/2}) \quad .$$

Also, the asymptotic variance of the maximum likelihood estimators $\hat{q}$ and $\hat{\beta}$ are related by

$$\text{var}(\hat{q}) = \frac{1}{4} \beta^{-3} \text{ var}(\hat{\beta}) \quad .$$

Parameterization B leads to some useful results. First, previous authors found that a large sample size was required when trying to discriminate between $\beta = 1$ and $\beta = 2$ or 3 . Prentice concluded that on a log scale these distributions are very similar and in fact, without a large sample size, $\beta = 1$ is difficult to discriminate from $\beta = \infty$. Furthermore, when $\hat{q} = 0$ ($\hat{\beta}$ close to infinity), maximum likelihood estimates from parameterization A cannot be obtained by solving the log likelihood equations. Using parameterization B, Prentice was able to obtain a log likelihood function which exists for all q and from which maximum likelihood estimates could be obtained given an adequate sample size. By simulation, Prentice showed that $\hat{q}$ converged faster to its asymptotic normal distribution than $\hat{\beta}$. An algorithm in SAS: PROC LIFEREG computes maximum likelihood estimates using parameterization B.

Application of Maximum Likelihood to the Nutrient Intake Data

The above algorithms (corresponding to parameterization A and parameterization B) were applied to data collected in 1985 by the USDA in the Continuing Survey of Food Intakes by Individuals (CFSII). The

survey collected daily dietary intakes from women between 19 and 50 years of age and from their preschool children. The data set used for our purposes was a subset of these data containing four days of dietary intakes for 785 women aged between 23 and 50 years who were responsible for meal planning within the household and who were not pregnant or lactating during the survey period. Six dietary components were analyzed including calcium, energy, iron, protein, vitamin A, and vitamin C. The data to which the algorithm was applied were predicted "pseudo" usual intakes generated from the original data using measurement error techniques [Nusser et al. (1990)]. This methodology was designed to adjust the observed intakes for the presence of measurement error in the dietary intake data.

The algorithm created by Nusser et al. (1990) involved several steps to produce pseudo usual intake values. First, the original daily intakes were transformed into normal space using a nonparametric transformation based on the inverse normal cumulative distribution function. In normal space, the daily intakes were assumed to follow a measurement error model, and normal theory was used to develop predictors of usual intakes in normal space for each individual. An inverse transformation was then applied to the predicted normal usual intakes to produce the set of pseudo usual intakes in the original space.

Applying the maximum likelihood estimation algorithm to the pseudo usual intake data for five of the nutrients produced the parameter estimates shown in Table 1. Using parameterization A, the algorithm failed to converge to the maximum likelihood parameter estimates for

vitamin A. Asymptotic standard errors were computed as the inverse of the estimated information matrix.

The asymptotic standard errors of the estimated coefficients are large, indicating the parameter estimates are relatively imprecise. This is due to the high correlation between the parameters. For all nutrients, the correlation between β and λ was between -0.99 and -1.00. Hager and Bain (1970) indicated that as β increases away from one, the asymptotic variance of $\hat{\beta}$ approaches infinity.

Table 1. Estimated Generalized Gamma parameters for each dietary component.

Dietary Component	$\hat{\theta}_{MLE}$	$\hat{\beta}_{MLE}$	$\hat{\lambda}_{MLE}$	Calculated log-likelihood
Calcium	35.84114 (51.61706)	9.65852 (5.19540)	0.81871 (0.22681)	-5327.88
Energy	320.46719 (243.85407)	7.75763 (3.73358)	1.32358 (0.33039)	-5811.81
Iron	0.35396 (0.72619)	16.73130 (11.8838)	0.84740 (0.30629)	-1912.46
Protein	20.84224 (10.31079)	5.88937 (2.46468)	1.66667 (0.36583)	-3218.55
Vitamin C	27.35708 (12.82631)	3.23953 (1.00431)	1.13405 (0.19121)	-3892.81

Table 2. Estimated Generalized Gamma parameters for each dietary component.

Dietary Component	$\hat{\alpha}_{MLE}$	$\hat{\sigma}_{MLE}$	$\hat{q}_{MLE}$	Calculated log-likelihood
Calcium	6.34976 (0.02182)	0.39289 (0.01037)	0.32491 (0.08405)	-5327.90
Energy	7.31722 (0.01480)	0.27135 (0.00719)	0.35623 (0.08138)	-5811.84
Iron	2.28698 (0.01476)	0.28834 (0.00741)	0.25092 (0.07291)	-1912.48
Protein	4.10782 (0.01329)	0.24566 (0.00669)	0.46814 (0.07951)	-3218.35
Vitamin A	8.27385 (0.03034)	0.53854 (0.01359)	0.01608 (0.08718)	-7119.67
Vitamin C	4.34508 (0.02841)	0.49006 (0.01412)	0.55409 (0.08848)	-3892.86

Using parameterization B, the LIFEREG procedure in SAS converged to the parameter estimates in Table 2. The estimates of Table 1 transformed into the parameterization of Table 2 are very close to those of Table 2. The largest difference occurred for iron, where β from parameterization A differs by 0.85 from the β calculated by parameterization B. Also, the calculated log likelihoods obtained by the two procedures are similar but not identical.

Note that for vitamin A, maximum likelihood estimates under parameterization A would be $\hat{\beta}_{MLE} = 3867.47853$, $\hat{\lambda}_{MLE} = 0.00048$ and $\hat{\theta}_{MLE} = e^{-1796.48}$. Thus, $\hat{\beta}_{MLE}$ is quite large, and as indicated by Prentice, this caused convergence problems for Vitamin A when the maximum likelihood estimation algorithm based on parameterization A was used.

It is of interest to test the fit of the hypothesized generalized gamma distributions for each nutrient. A chi-square goodness-of-fit statistic, using twenty-five mutually exclusive intervals over the range of the pseudo usual intake data, was used as a test statistic. Observed frequencies for the pseudo usual intake data and the expected frequencies from the hypothesized distributions were computed and the test statistics, based on 21 degrees of freedom, are presented in Table 3 for each of the six dietary components. Tests of size 0.05 indicate the generalized gamma provides a satisfactory fit for all six of the dietary components. A plot of the hypothesized generalized gamma and empirical cumulative distribution function for each nutrient is included in Appendix A.

Testing Hypotheses with the Maximum Likelihood Estimation Algorithm

Of interest is whether or not the Weibull or gamma distributions would provide a satisfactory fit for the pseudo usual intake data. Tests of these hypotheses can be constructed using the likelihood ratio test.

Table 3. Goodness of Fit test for Generalized Gamma Distribution.

Component	χ^2
Calcium	27.3
Energy	22.5
Iron	32.5
Protein	27.4
Vitamin A	24.3
Vitamin C	15.6

The $\alpha = .05$ point of the chi-square distribution with 21 df is 32.7.

Table 4. Estimated Weibull parameters for each dietary component.

Dietary Component	$\hat{\theta}_{MLE}$	$\hat{\lambda}_{MLE}$
Calcium	652.26217 (9.13450)	2.65716 (0.07395)
Energy	1641.63778 (15.82403)	3.86013 (0.10724)
Iron	10.94775 (0.11705)	3.47975 (0.09684)
Protein	64.85339 (0.56710)	4.25585 (0.11832)
Vitamin A	5109.62014 (101.22994)	1.87827 (0.05227)
Vitamin C	85.95530 (1.47753)	2.16479 (0.06025)

Testing $\beta = 1$. The density function for the Weibull distribution is

$$f_W(y) = \theta^{-1}\lambda(\theta^{-1}y)^{\lambda-1}\exp[(-\theta^{-1}y)^{\lambda}] ,$$

which is the generalized gamma density with $\beta = 1$. By constraining the maximum likelihood estimator algorithm to estimate the parameters of the Weibull distribution, the parameter estimates for the six nutrients listed in Table 4 were obtained. Note that when β is set equal to one, the standard errors of the parameter estimates decrease dramatically compared to those computed under the generalized gamma assumption. To test whether the Weibull is a suitable family of distributions for each nutrient, likelihood ratio tests were constructed. By the asymptotic properties of the likelihood ratio test, $2(\log \text{likelihood}_{WEIB} - \log \text{likelihood}_{GGD})$ is asymptotically distributed

Table 5. Likelihood ratio test for Weibull null hypothesis against Generalized Gamma.

Component	χ^2
Calcium	63.2
Energy	58.0
Iron	100.2
Protein	42.7
Vitamin A	118.5
Vitamin C	24.1

The $\alpha = .05$ point of the chi-square distribution with 1 df is 3.84.

as chi-square with one degree of freedom under the null hypothesis. The test statistics are presented in Table 5. Tests of size 0.05 indicate that the Weibull family does not adequately fit the distribution of any of the dietary components analyzed. That is, the generalized gamma hypothesis dominates the Weibull hypothesis in the likelihood ratio test.

Asymptotic chi-square goodness-of-fit tests for the hypothesized Weibull distributions were computed and are listed in Table 6. These statistics also indicate that the hypothesized Weibull distribution does not accurately describe the data for five of the dietary components, but is adequate for vitamin C. Plots of the hypothesized Weibull and empirical cumulative distribution functions are included in Appendix A.

Testing $\lambda = 1$. The density function for the gamma distribution can be written

$$f_G(y) = [\theta\Gamma(\beta)]^{-1}(\theta^{-1}y)^{\beta-1}\exp[-\theta^{-1}y] \quad ,$$

Table 6. Goodness of fit test for Weibull Distribution.

Component	χ^2
Calcium	54.6
Energy	62.5
Iron	84.5
Protein	37.3
Vitamin A	115.7
Vitamin C	31.5

The $\alpha = 0.05$ point of the chi-square distribution with 22 df is 33.9.

which is the generalized gamma distribution with $\lambda = 1$. Estimates of the gamma distribution parameters for the six dietary components are given in Table 7. Note that standard errors are again much lower than those for the generalized gamma distribution.

Again, $2(\log \text{likelihood}_{GAM} - \log \text{likelihood}_{GGD})$ is asymptotically chi-square with one degree of freedom under the null hypothesis. The

Table 7. Estimated Gamma parameters for each dietary component.

Dietary Component	$\hat{\theta}_{MLE}$	$\hat{\beta}_{MLE}$
Calcium	88.20452 (4.51406)	6.57410 (0.32376)
Energy	111.68941 (5.67370)	13.33135 (0.66456)
Iron	0.81781 (0.04157)	12.10267 (0.60258)
Protein	3.79538 (0.19262)	15.59343 (0.77873)
Vitamin A	1246.43305 (64.61994)	3.61820 (0.17487)
Vitamin C	18.62355 (0.96226)	4.07795 (0.19799)

likelihood ratio test statistic for each nutrient was computed and is presented in Table 8. The likelihood ratio test at size 0.05 indicates the gamma family cannot be rejected as an adequate family to describe the distribution for the dietary components calcium, energy, iron and vitamin C.

An approximate chi-square goodness-of-fit test for the hypothesized gamma distributions were computed and are listed in Table 9. Corresponding plots of the hypothesized gamma and empirical cumulative distribution function for each nutrient are included in Appendix A. Chi-square tests of size 0.05 indicate that the hypothesized gamma distributions is satisfactory in describing the distribution of the data for five dietary components, vitamin A excluded.

Table 8. Likelihood ratio test χ^2 for gamma null hypothesis against Generalized Gamma.

Component	χ^2
Calcium	0.6
Energy	1.0
Iron	0.2
Protein	6.8
Vitamin A	34.0
Vitamin C	0.5

The $\alpha = .05$ point of the chi-square distribution with 1 df is 3.84.

Table 9. Approximate goodness-of-fit test χ^2 for each dietary component gamma parameter estimates.

Component	χ^2
Calcium	28.4
Energy	21.4
Iron	30.0
Protein	30.2
Vitamin A	54.5
Vitamin C	22.7

The $\alpha = .05$ point of the chi-square distribution with 22 df is 33.9.

Using the Estimated Usual Intake Distributions to Estimate the Prevalence of Nutritional Inadequacy

Estimated usual intake distributions can be used to evaluate the nutritional status of a population of individuals. The assessment of dietary status within a population typically involves comparison of observed dietary intakes with a measure of the requirement for a particular nutrient or food component (NRC, 1986). One common method of comparison relies on a fixed point cut-off requirement level. The recommended daily allowance for a dietary component, established by the United States Department of Agriculture Food and Nutrition Board, is an example of such a cut-off point. The cut-off method utilizes a standard requirement level as the criterion value, where individuals with intakes below this standard are said to be at nutritional risk. Since the recommended daily allowance levels are set sufficiently high to meet the known nutritional needs of nearly all healthy persons, a cut-off point often is defined to be a proportion of the recommended daily allowance. The proportion used by researchers and policy makers differs

across nutrients and studies. The choice of a proportion, k, is a matter of judgment, though the cut-off proportion is usually chosen between 0.5 and 0.8. Any estimate of an at risk percentage is very sensitive to the value of the recommended daily allowance proportion k. As this proportion decreases, the percentage of the population deemed at risk declines.

Given a maximum likelihood estimate of the usual intake distribution and a proportion k, estimates of the percentage of the population of women aged 23-50 who are at nutritional risk for a given dietary component can be obtained. The cumulative distribution function for the generalized gamma, defined in equation (2), is used to calculate the proportion of the population whose intakes fall below a specified level y. For the six dietary components and for $k = 0.5, 0.65, 0.8$, the estimates of women aged 23-50 at nutritional risk for a given dietary component using the generalized gamma models are presented in Table 10.

Table 10. The estimated percentage at-risk (and approximate standard errors) for each dietary component using generalized gamma parameter estimates.

Component	Percentage at-risk for criteria k(RDA) as the cut-off point			
		k		
	RDA	.5	.65	.8
Calcium (mg)	800	22.42 (1.23)	44.76 (1.56)	65.50 (1.47)
Energy (kcal)	2,000	10.36 (0.88)	34.30 (1.44)	63.47 (1.46)
Iron (mg)	18	41.18 (1.48)	75.70 (1.24)	93.02 (0.74)
Protein (g)	44	0.15 (0.10)	1.05 (0.26)	4.17 (0.50)
Vitamin A (IU)	2666	2.35 (0.85)	6.63 (0.81)	13.11 (1.17)
Vitamin C (mg)	60	7.52 (0.78)	14.98 (1.08)	24.35 (1.26)

Since the distribution function is nonlinear in the parameters, the standard errors for the percentage estimates are approximated using Taylor's theorem. Denote the estimated cumulative distribution function for the generalized gamma as $F(y; \hat{\theta}, \hat{\beta}, \hat{\lambda})$. By Taylor's theorem we have

$$F(y; \hat{\theta}, \hat{\beta}, \hat{\lambda}) \doteq F(y; \theta, \beta, \lambda) + \frac{dF(y; \theta, \beta, \lambda)}{d\theta} (\theta - \hat{\theta})$$

$$+ \frac{dF(y; \theta, \beta, \lambda)}{d\beta} (\beta - \hat{\beta}) + \frac{dF(y; \theta, \beta, \lambda)}{d\lambda} (\lambda - \hat{\lambda}) + O_p(n^{-1}) ,$$

where (θ, β, λ) are the true parameters. Thus,

Table 11. The estimated percentage at risk (and approximate standard errors) for each dietary component using gamma parameter estimates.

Component	Percentage at risk for criteria k(RDA) as the cut-off point			
		k		
	RDA	.5	.65	.8
Calcium (mg)	800	22.29 (1.21)	44.32 (1.41)	65.05 (1.37)
Energy (kcal)	2,000	10.22 (0.86)	34.75 (1.37)	63.93 (1.38)
Iron (mg)	18	40.95 (1.40)	75.59 (1.25)	93.11 (0.69)
Protein (g)	44	0.05 (0.02)	0.67 (0.13)	3.59 (0.46)
Vitamin A (IU)	2666	4.09 (0.50)	8.33 (0.77)	13.97 (1.00)
Vitamin C (mg)	60	7.36 (0.72)	14.96 (1.03)	24.53 (1.25)

$$\mathrm{Var}\left[F(y;\ \hat{\theta},\ \hat{\beta},\ \hat{\lambda}) - F(y;\ \theta,\ \lambda,\ \beta)\right] \doteq V\left[\frac{dF}{d\rho}\ (\hat{\rho} - \rho)\right] ,$$

where $\rho = (\theta,\ \beta,\ \lambda)$. That is,

$$\mathrm{Var}\left[F(y;\ \hat{\theta},\ \hat{\beta},\ \hat{\lambda}) - F(y;\ \theta,\ \lambda,\ \beta)\right] \doteq \frac{dF(y,\ \rho)}{d\rho}\ \hat{V}(\hat{\rho})\ \frac{dF(y,\ \rho)}{d\rho} ,$$

where $\hat{V}(\hat{\rho})$ is the estimated covariance matrix of the parameter estimates. Since the values of the partial derivatives at the true parameters are not known, the parameter estimates are used to

approximate the partial derivatives as well. Hence, the variance of the percentage estimates can be approximated by

$$\hat{\text{Var}}[F(y;\ \hat{\theta},\ \hat{\beta},\ \hat{\lambda})] \doteq [\frac{dF}{d\theta},\ \frac{dF}{d\beta},\ \frac{dF}{d\lambda}]\ \hat{\mathbf{V}}(\hat{\theta},\ \hat{\beta},\ \hat{\lambda})\ [\frac{dF}{d\theta},\ \frac{dF}{d\beta},\ \frac{dF}{d\lambda}]' \ ,$$

where $[\frac{dF}{d\theta},\ \frac{dF}{d\beta},\ \frac{dF}{d\lambda}]$ are evaluated at the parameter estimates and $\hat{\mathbf{V}}(\hat{\theta},\ \hat{\beta},\ \hat{\lambda})$ is the estimated covariance matrix of the parameter estimates.

A relatively high percentage of the population is estimated to have nutritional deficiencies in calicum, energy and iron, and a low percentage is estimated to be at risk with respect to protein and Vitamin A.

The percentage of women aged 23-50 estimated to be at nutritional risk, based on the gamma models, are presented in Table 11. The percentage at-risk estimates are similar under the generalized gamma and gamma models, for all dietary components excluding Vitamin A. Recall that the gamma model was found not to be a satisfactory model for the Vitamin A data. Also, although the standard errors for the generalized gamma parameter estimates are large, the standard errors of the percentages at-risk under the generalized gamma models are only slightly larger than the standard errors computed under the gamma model.

Conclusion

In this paper, estimation of the generalized gamma distribution has been discussed. An algorithm has been presented which generates starting values from the data, and uses these starting values to obtain maximum likelihood estimators for the generalized gamma parameters.

Maximum likelihood estimates of the parameters of the distribution of usual intakes were obtained using two parameterizations. Likelihood ratio tests were used to test whether the distribution of usual intakes for selected dietary components could be reasonably described by the less complex Weibull and gamma families. Finally, using the maximum likelihood estimators of the generalized gamma parameters, the percentage of women aged 23-50 at nutritional risk was estimated using the cut-off method at different proportions of the recommended daily allowances for the dietary components analyzed in this study.

This study indicates that for the 1985-1986 CSFII data, the less complex gamma distribution provides an adequate fit for five of the nutrients, vitamin A excluded. Based on the chi-square goodness of fit tests, the generalized gamma family appears to accurately represent all six nutritional component usual intake distributions. Although the standard errors of the parameter estimates are large, the standard errors of the statistic of interest, percentage at-risk, are reasonable and allow for substantive conculsions regarding nutritional inadequacy to be made.

REFERENCES

Bernardo, J.M. 1976. "Psi (Digamma) Function." Applied Statistics 25, p. 315-17.

Hager, H.W., and L.J. Bain. 1970. "Inferential Procedures for the Generalized Gamma Distribution." Journal of the American Statistical Association 65, p. 1601-9.

Harter, H.L. 1967. "Maximum-Likelihood Estimation of the Parameters of a Four-Parameter Generalized Gamma Population from Complete and Censored Samples. Technometrics 9, p. 159-65.

IMSL Inc. 1987. IMSL Math/Library: Fortran Subroutines for Mathematical Applications; Vol. 3., Houston, Texas.

National Research Council. 1986. Nutrient Adequacy: Assessment Using Food Consumption Surveys, Washington, D.C.: National Academy Press.

Nusser, S.M., G.E. Battese, and W.A. Fuller. 1988. "Estimation of the Distribution of Usual Intakes for Selected Dietary Components Using Data from the 1985-86 Continuing Survey of Food Intake by Individuals," report submitted to the Human Nutrition Information Service, U.S. Department of Agriculture, p. 56.

Parr, V.B., and J.T. Webster. 1965. "A Method for Discriminating Between Failure Density Functions Used in Reliability Predictions." Techometrics 61, p. 539.

Prentice, R.L. 1974. "A Log Gamma Model and Its Maximum Likelihood Estimation." Biometrika 61, p. 539.

SAS Institute Inc. 1985. SAS User's Guide: Basics. 1985 Ed., Cary, North Carolina.

SAS Institute Inc. 1985. SAS/IML User's Guide, 1985 Ed., Cary, North Carolina.