DISTRIBUTIONS AND TRANSFORMATIONS FOR FAMILY EXPENDITURES

Stuart Scott and Daniel J. Rope, Bureau of Labor Statistics
2 Massachusetts Avenue, NE, Rm 4915, Washington, D.C. 20212

1. Introduction

The aims of this paper are to study family expenditure distributions to determine

(1) whether transformations can improve the normal approximation for inference purposes,

(2) whether selected probability models provide adequate fits to the data.

The data are from the U. S. Consumer Expenditure Quarterly Interview Survey for 1984 and 1985. The survey, conducted on a continuing basis since 1980, samples 5000 consumer units each quarter, as described in BLS's Handbook of Methods (1992). The data are collected for consumer units, defined in the handbook, but in this paper we will more informally speak of families or households. Usually, the concepts coincide.

This survey is the source of base period expenditures for the U.S. Consumer Price Index. Additional analytic uses include spending comparisons, either across time or demographic groups (e.g. Jacobs, Shipp, and Brown, 1989, and Wagner and Soberon-Ferrer, 1990). Regression models have been used to estimate income elasticity of expenditures for certain categories (Gieseman and Moulton, 1987). The family budget program, described in Watts (1980), but discontinued in 1981, estimated typical expenditures for two family types, (1) a husband, wife, and two children and (2) a retired couple. Researchers frequently use transformations in carrying out analyses of expenditure data. Wagner and Soberon-Ferrer take log's of expenditures, and Gieseman and Moulton use square roots of food expenditures and family income.

Expenditure distributions are usually skewed right. For instance, within Women's Apparel, there are frequent purchases of smaller items, such as hosiery or shirts, and scattered purchases of more costly items, such as suits or overcoats. "Lifetime" or "failure time" distributions share this feature of positive skewness, and have received detailed attention from statisticians. Lawless (1982) in his text *Statistical Models and Methods for Lifetime Data*, gives a unified treatment of common probability models for lifetimes.

We will select Box-Cox transformations for our data using both the classic maximum likelihood method and alternative approaches suggested by Hernandez and Johnson (1980), and Lin and Vonesh (1989).

2. Expenditure Distributions

Following Lawless, we use the generalized gamma distribution as a unified means of carrying out probability modeling of expenditure distributions. The generalized gamma density is defined as

$$f_{GG}(x) = \frac{1}{\Gamma(k)} \cdot \frac{\beta}{\alpha^{\beta k}} x^{\beta k - 1} e^{-(x/\alpha)^{\beta}}, \ x > 0.$$

α is a simple scale parameter, and β and k are shape parameters. Special cases are

Weibull:	k =1
gamma:	β =1
exponential:	$k = \beta$ =1.

The lognormal distribution is a limiting case of the gamma.

Figure 1 shows possible shapes of these distributions. The lognormal rises quickly to a peak, drops quickly, and then declines fairly slowly, and it has the thickest tail among these distributions. The shape of the generalized gamma depends on the product $\beta\, k$.

$$\beta k < 1 \Rightarrow f_{GG}(x) \to \infty \text{ as } x \to 0$$

$$\beta k = 1 \Rightarrow f_{GG}(x) \text{ mode at } x = 0$$

$$\beta k > 1 \Rightarrow f_{GG}(x) \text{ unimodal, (unique positive mode)}$$

Examples in the figure are the Weibull with β =.5, the exponential, and the gamma with k =2, respectively.

Empirical distributions have been examined for two years, four expenditure categories (cf. Table 1), and five income classes. In thousands of dollars, the income classes are [0,10), [10,20), [20,30), and [30,∞), plus a class of incomplete income reporters. Each observation represents total household spending in the category for a month. All these results are for the positive part of the distribution. The full distribution consists of a spike at zero for families making no purchases in the category, plus the spending distribution. Table 1 shows the overall reporting rates for positive spending for the four categories. Even for the broad categories Home Furnishings & Equipment and Women's Apparel, just over one-third of the families report expenditures. The forty spending distributions are each fitted by maximum likelihood to the four basic distributions described above via an adapted program by Jacqueline Kent (cf. Quesenberry and Kent, 1982). Lawless (1982, Chapter 5) refers to

Table 1. Percent Households Reporting Expenditures

Category	Percent
Home Furnishings& Equipment	38
Home Appliances	3
Women's Apparel	36
New Cars & Trucks	1

computing difficulties in earlier years in finding maximum liklihood solutions of the generalized gamma and, based on a reparameterization, shows how the maximum may be found from solving a single interative equation. Empirical and fitted distributions are graphed, and chi-square goodness-of-fit testing carried out. (Relative survey weights are incorporated into the empirical distributions. Since we will not be interpreting p-values too strictly, we have not made adjustments for the complex survey design).

Let us consider in detail Women's Apparel, Income class 1, which has 4396 observations for 1984. The right skew is evident in the histogram in Figure 2. Even though over 80% of the distribution lies below \$100, values range to \$3451. The mean, \$63, is more than double the median of \$30 and Fisher's skewness coefficient is +11. The histogram shows considerable roughness in the data. Especially noteworthy are large spikes, occurring at \$10 intervals. For instance, 67 reported expenditures are \$50, but only 17 reported expenditures are \$49 and 18 are \$51. Respondents tend to round off. Also, in part, this reflects pricing strategies. Retailers are more likely to price an item at \$19.99 than \$19.00.

Among the four basic models, the largest p-value (.02) comes from the chi-square test for the lognormal distribution. This rather poor fit is the second best for the lognormal out of the ten income class-year combinations. The curvature of the lognormal fits these data better than the other basic distributions (cf. Figure 3). Still, it tends to be too high between 5 and 25, and too thick in the tail. A distinct improvement in fit is achieved with the generalized gamma distribution with b =.25 and k =9.5. The p-value is .38, suggesting a comfortable acceptance of the model. The generalized gamma achieves better p-values than the lognormal for all ten income-year combinations, and in six cases gives acceptable p-values (cf. Table 2).

Table 2 shows which income ▪ year classes have adequate (denoted '+') and inadequate (denoted '0') fits. The roughness of the data makes model-fitting difficult. We did not try to account for the spikes, but did try to minimize their effects by a careful choice of cells. A considerable part of the difficulty is the inherent diversity in these distributions. The CE Survey encompasses all types of households from all

Table 2. Best-Fitting Distributions (1984/1985)

Income Class	Home Furnishings	Home Appliances	Women's Apparel	New Cars and Trucks
<$10,000	LN/LN (0/0)	EE/EE (+/+)	GG/GG (+/+)	WB/WB (+/+)
$10K-20K	LN/LN (0/0)	EE/EE (.02/+)	GG/GG (+/0)	GM/GM (+/+)
$20K-30K	LN/LN (0/0)	EE/EE (0/0)	GG/GG (.06/+)	WB/WB (+/0)
$30K and +	LN/LN (0/0)	WB/WB (0/.02)	GG/GG (0/0)	GM/WB (0/.07)
Incomplete Reporter	LN/LN (0/0)	EE/EE (+/+)	GG/GG (+/0)	WB/GM (0/.08)

EE = Exponential WB = Weibull GM = Gamma GG = Generalized Gamma LN = Lognormal
('+' denotes p-value > .10, '0' denotes p-value < .01, Other entries are best p-values obtained.)

parts of the country. Moreover, even within a family, across months, spending shifts in terms of categories and amounts. Households with the largest incomes ($30,000 or more), form the largest class in the population and also the hardest class to fit. Women's Apparel purchases range from pairs of socks to fur coats. Home Furnishings and Equipment has no acceptable fits, while the more homogeneous subcategory Home Appliances has several. While less diverse, perhaps, than the other categories, New Cars and Trucks has comparatively small sample sizes. The success in fitting most classes is due in part to the small number of cells used in the chi-square testing.

Results for the best-fitting distribution (shown in Table 2) are fairly consistent across income class. Parameter values do not differ greatly with income. Spending levels do not increase with income as much as expected. As just discussed, the generalized gamma is chosen for Women's Apparel. Surprisingly, the exponential gives an acceptable fit for Major Appliances. Either the Weibull (with beta's around 2) or the gamma (with k's in the 2.5-5 range) is selected for New Cars and Trucks. For Homefurnishings and Equipment, the lognormal outperforms all the other distributions, including the generalized gamma.

3. Transformations

The probability models in the previous section may be used for description, but usually for inference it is desirable to have approximate normality. The classic paper by Box and Cox (1964) gives a method for selecting a transformation to improve the normal approximation, both for the basic distribution and for the mean derived from it. The Box-Cox family of transformations is

$$X^{(\lambda)} = \begin{cases} \dfrac{x^{\lambda} - 1}{\lambda}, & \lambda \neq 0, \\ \log x, & \lambda = 0. \end{cases}$$

(The particular form of the power transformation is chosen for continuity at λ=0. That is,

$X^{(\lambda)} \to \log X$ as $\lambda \to 0$.) The basic approach is to compute the likelihood function under a normal model for a range of λ, and pick out the λ with the largest likelihood. Because the likelihood function is flat around the maximum, there is no need to use an exact value. It seems preferable to restrict the transformation to relatively simple values, such as 0, 1/2, 1/4, etc.

For Homefurnishings and Women's Apparel, the optimal values range between -.075 and .05 for all income classes, so the log transformation is appropriate. Table 3 contains the optimal λ for Home Appliances. With values ranging from 0.02 to .25, and some variation across year, one compromise solution is to select a sixth root for all size classes. (The average optimal λ is 0.16). For New Cars and Trucks, Table 3 shows a much greater range of values, from 0.3 to 1.0. There is no monotonic trend with income, however, and there seems to be about as much variation within income class as across. If one wishes to use a single λ for this category, a rough compromise would be λ=3/4.

Table 3. Optimal λ from MLE (1984/1985)

Income Class	Home Appliances	New Cars and Trucks
< $10K	.015 / .175	1.020 / .475
$10K-$20K	.125 / .175	.850 / .615
$20K-$30K	.200 / .175	1.015 / .980
$30K and over	.250 / .150	.300 / .750
Incomplete Reporter	.130 / .150	.700 / .415

Can model selection contribute to transformation selection? Hernandez and Johnson (1980) utilize a distance measure to find the optimal λ for transforming a known distribution to achieve approximate normality. Their results for the gamma family extend to the generalized gamma family.

Given two probability densities *g* and *h*, the Kullback-Leibler information number is defined as

$$I(g,h) = \int g(x)\cdot \log \frac{g(x)}{h(x)}\,dx.$$

It serves as a distance measure between *g* and *h*. In particular, if *g=h,* then I is 0. If $X \sim g$, $\mu = E(X)$, $\sigma^2 = Var(X)$, then $I(g,\phi_{\mu,\sigma^2})$ measures the closeness of *g* to the normal distribution with the same mean and variance. A general formula for G(λ) the minimum $I(f_\lambda, \phi_{\mu,\sigma^2})$ over μ and σ^2 is (with $g = f_\lambda$)

$$G(\lambda) = \frac{1}{2}\left[\log(2\pi)+1\right] + E_g\left\{\log\left[g(X)\right]\right\}$$
$$+ (1-\lambda)E_g\left[\log(X)\right] + \frac{1}{2}\log\left[Var_g(X^{(\lambda)})\right].$$

For the generalized gamma function, this reduces to

$$G(\lambda) = \frac{1}{2}\left[\log(2\pi)+1\right] + \log\beta - 2\log\Gamma(k) + (k - \frac{\lambda}{\beta})\psi(k)$$
$$-k + \frac{1}{2}\log\left[\frac{\Gamma(k+\frac{2\lambda}{\beta})\Gamma(k) - \Gamma^2(k+\frac{\lambda}{\beta})}{\lambda^2}\right],\ \lambda \neq 0,$$

and

$$G(0) = \frac{1}{2}\left[\log(2\pi)+1\right] - \log\Gamma(k) + k\psi(k) - k + \frac{1}{2}\log\psi'(k),$$

where $\psi = \psi(k) = \frac{d}{dk}\log\Gamma(k)$ is the digamma function.

From the function G(λ), one can determine the best value λ^*.

EXAMPLE. Gamma distribution with k=2.

The optimal λ is .30, with G(.30)=.0005. This compares with

G(1) = .188
G(.5) = .018
G(0) = .060.

The Kullback-Leibler information number is a distance function, but not one that we are accustomed to. In relative terms at least, the closeness to normality is greatly improved by the optimal transformation and also by the cube root, G(1/3) = .001.

For the entire gamma family, Figure 4 indicates how much transformations matter by plotting G(λ)=G(λ,k) as a function of the shape parameter k for several fixed values of λ. Here is the graph with λ=1, .5, .33, .25, and 0. Asymptotically, λ^*=1/3, as shown by Hernandez and Johnson. Visually, at least, it appears that most of the gain comes for values k<1, although large percentage gains are achieved for larger k.

Looking at Figure 4 in terms of λ, λ=1 has the largest values for all k. The log transformation outperforms some of the root transformations near 0, but beyond k=.6 all three root transformations are superior.

Table 4. Optimal l from MLE vs. Models, 1985 data

	Box-Cox MLE	Generalized Gamma Model	Best Other Model
Income Class 1 (0-$10K)			
Home Furnishings	-.075	.075	0
Home Appliances	.175	.200	.275
Women's Apparel	.050	.075	0
New Cars & Trucks	.475	.525	.800
Income Class 4 ($30 and over)			
Home Furnishings	-.050	.075	0
Home Appliances	.150	.150	.275
Women's Apparel	.050	.075	0
New Cars & Trucks	.750	.700	.800

We now apply these results to picking transformations for expenditure distributions. Taking the best-fitting model for each family, we can compute G(l) as a function of l, and pick an optimal value. Table 4 shows l^* from direct maximum likelihood, the best-fitting basic distribution, and the generalized gamma for 1985 data. Recall that Income class 1 had adequate fits for all categories except Homefurnishings. Income class 4 was the hardest fit, with adequate fit only for New Cars & Trucks.

For Women's Apparel, the optimal l is .075, based on the generalized gamma, the only model giving an adequate fit. Maximum likelihood gives virtually the same result, l =.05. Either value suggests that the log transformation is a near-optimal choice. The values from the exponential, gamma, and Weibull are .275, .250, and .225, and in all cases the distance from normality is much greater than for the lognormal. The generalized gamma results are very close to direct MLE in all cases. The largest difference is .05, and in one of these cases, New Cars & Trucks, Income class 1, both suggest a square root transformation.

Using the best-fitting basic distribution is not as effective. For Homefurnishings and Women's Apparel, where the lognormal fits best, there is agreement with MLE. For Home Appliances, the exponential and Weibull models are used, and give too high a value of l, but using the value implied by these models, l =1/4, probably works fairly well. For New Cars & Trucks, the value is quite a bit off for Income class 1, but close for Income class 4.

Lin and Vonesh (1989) propose setting up a nonlinear regression in l to select the transformation. A prime reason offered for the method is convenience, in the sense that many users may have access to nonlinear regression software, but not an existing MLE program. We did not find this method convenient, due to some problems with convergence. There seem to be very high correlations among the estimated parameters, raising concerns about stability in estimation.

4. Summary

With respect to transformation selection, a log transformation is appropriate for Homefurnishings and Women's Apparel. A compromise choice for Home Appliances is a one-sixth root. For New Cars & Trucks, the choice might range from a square root to no transformation, depending on the income class.

Fitting distributions is difficult for these data. There is roughness in the data in the form of spikes at even values. Except perhaps for New Cars & Trucks, items in the expenditure categories are quite diverse. Also, the reporting households are diverse; the income breakdown does not really seem to help that much. Even so, adequate fits are obtained in about half the cases. There was not much difference in the choice of model by income class. Here are the distributions selected:

Homefurnishings	lognormal
Home Appliances	exponential
Women's Apparel	generalized gamma
New Cars & Trucks	Weibull or gamma

Is modeling the distributions useful? If one is simply seeking a transformation to improve normality for an expenditure distribution, basic MLE is simple and works well. Modeling, however, in our opinion, may be helpful in identifying patterns in expenditure distributions. Rather than repeatedly computing MLE's, one may find similar behavior for certain expenditures, even across demographic groups. The Hernandez-Johnson statistics assess distance from normality.

The generalized gamma has been useful in several ways: it provides a unified framework for studying all the probability models in this paper. It is useful in its own right. Its flexibility gives a good fit to Women's Apparel, where the other distributions fail. Finally, for transformation selection with the Hernandez-Johnson formulas, it performs better than the simple distributions, and agrees quite well with the MLE approach.

References

Box, G.E.P. and Cox, D.R.(1964), "An Analysis of Transformations," *Journal of the Royal Statistical Society*, Ser. B 26, 211-243, discussion, 244-252

Gieseman, Raymond W. and Moulton, Brent R. (1987), "Estimates of the Income Elasticity of Expenditures for Food in 1985," *Proceedings of the ASA Business and Economic Statistics Section*, 529-534

Hawkins, Douglas M. and Wixley, R. A. J. (1986), "A Note on the Transformation of Chi-Squared Variables to Normality," *American Statistician* 40, 296-298

Hernandez, Fabian and Johnson, Richard A. (1980), "The Large-Sample Behavior of Transformations to Normality," *Journal of the American Statistical Association* 75, 855-861

Jacobs, Eva, Shipp, Stephanie, and Brown, Gregory (1989), "Families of Working Wives Spending More on Services and Nondurables," *Monthly Labor Review*, 15-23

Lawless, J. F. (1982), *Statistical Models and Methods for Lifetime Data*, Wiley:New York

Lin, Lawrence I-kuei and Vonesh, Edward F. (1989), "An Empirical Nonlinear Data-Fitting Approach for Transforming Data to Normality," *American Statistician*, 237-243

Quesenberry, C. P. and Kent, Jacqueline (1982), *Technometrics* 24, 59-65

Wagner, Janet and Soberon-Ferrer, Horacio (1990), "The Effect of Ethnicity on Selected Household Expenditures," *Social Science Journal* 27, 181-198

Watts, Harold W. (1980), "Special Panel Suggests Changes in BLS Family Budget Program," *Monthly Labor Review*, 3-10

Acknowledgments

Thanks to Sharon Krieger and Karen Fung for earlier work on the project, to Sylvia Johnson-Herring for providing the data, to Stephanie Shipp and Adriana Silberstein for background information, and to Paul Hsen for overall support.