

**A COMPARISON OF METHODS OF ESTIMATION OF PARAMETERS  
OF TUKEY'S  $gh$  FAMILY OF DISTRIBUTIONS**

**M. Mahbubul A. Majumder**

Department of Statistics, Iowa State University  
Ames, Iowa 50010  
and

**M. Masoom Ali\***

Department of Mathematical Sciences, Ball State University  
Muncie, IN 47306  
Email: mali@bsu.edu

**ABSTRACT**

The  $gh$  family of distributions proposed by Tukey (1977) based on a transformation of the standard normal variable does not have any explicit mathematical form. Thus the study of this distribution requires extensive numerical computations. In this paper we study the numerical methods of estimating the parameters  $g$  and  $h$ . We compare the three methods namely quantile method, method of moments and maximum likelihood method by using simulation technique. We have found that maximum likelihood method is more efficient than the other two methods though it requires much more computational load.

**1. INTRODUCTION**

In practical applications the underlying distributions from which the data actually arise do not always match the assumed distributions. In most cases, the distributions of empirical data are different than the assumed distributions. Now a days, the large varieties of data have kept the statisticians busy in quest for new families of distributions to describe the data in general. But it is not an easy task to obtain a distribution which best fit the empirical data set.

To meet this purpose, Tukey (1977) introduced a family of distributions called  $g$ -and- $h$  family ( $gh$  family) based on a transformation of the standard normal variable. Let  $Z$  denote a standard normal variable, then the  $gh$  distribution of a univariate normal random variable  $Y_{gh}$  is defined through the following transformation of  $Z$

$$Y_{gh}(Z) = \mu + \sigma \frac{e^{gZ} - 1}{g} e^{hZ^2/2},$$

where,  $\mu$  is the location parameter,  $\sigma (> 0)$  is the scale parameter and  $g$  and  $h$  are the scalars that govern the skewness and elongation of  $Y_{gh}$ , respectively. The density of  $gh$

---

\* *Dr. M. Masoom Ali is George and Frances Ball Distinguished Professor Emeritus of Statistics and Professor Emeritus of Mathematical Sciences.*

distribution can only be expressed as an implicit function. Thus it requires numerical computation to obtain the estimates of  $\mu$ ,  $\sigma$ ,  $g$  and  $h$ . The  $gh$  family of distributions was extensively studied by Hoaglin (1985) and Martinez and Iglewicz (1984). Due to its appealing attributes in shape it has been getting popular for simulation studies. He and Raghunathan (2006) have used this distribution for multiple imputations. Despite its complex mathematical form, percentage points of the density function can be obtained numerically using various computer software packages. Hoaglin (1985) and Martinez and Iglewicz (1984) studied the properties of this family using computer packages.

The most important and useful characteristic of the  $gh$  family of distributions is that this family includes several known theoretical probability distributions. Table 1 shows the list of these distributions.

**Table 1:**  
**Values of  $g$  and  $h$  for Some Distributions**

Distributions	$g$	$h$
Normal	0	0
Log-Normal	1	0
Cauchy	0	0.97
$t$ distribution with $df=10$	0	.058
Uniform	0	-0.244
$\chi^2$ distribution with $df=4$	0.502	-0.046
Exponential	0.76	-0.098

We developed an algorithm to solve the equations to estimate the parameters  $g$  and  $h$  using the method of moments. We have found that the estimated parameters obtained using this algorithm is as good as that obtained by quantile method we have found that if we solve the equations by the method of moments using our algorithm, it reduces lot of computations.

To obtain maximum likelihood estimates of  $g$  and  $h$  it requires a heavy computational load. Even with the high speed Pentium-IV PC it takes a long time to obtain the solutions. In our study we obtained the maximum likelihood estimates of  $g$  and  $h$  for sample of size 300. For this we used Personal computer with Pentium-IV processor having 1.2 GHz of processing power and 512 MB of RAM. For quantile method and method of moments we used the same machine.

We simulated data from  $gh$  distribution. In our study of the performance of the estimation methods, we used simulation technique to estimate the standard errors of the estimates. We compared all the three methods having the standard error from the simulated samples. It is observed that the methods are equally good in respect of standard error though method of moments appeared to require less computation.

## 2. ESTIMATION OF PARAMETERS $g$ AND $h$

### 2.1 Quantile Method

The quantile method is commonly used for estimating the parameters  $g$  and  $h$ . Hoaglin (1985) showed how the method works. The idea is to estimate the parameter  $g$  first directly from the quantiles. That is to estimate  $g$  for which the  $p$ th and  $(1-p)$ th quantiles of the data fit the distribution exactly. Thus for each  $p$  we will estimate a value for  $g$ . So we can treat this estimate as a function of  $p$  namely  $g_p$ . Hoaglin showed that for each  $p$ th quantile of the data  $y_p$  the estimate of  $g_p$  is

$$g_p = -\frac{1}{z_p} \log_e \frac{y_{1-p} - y_{.5}}{y_{.5} - y_p},$$

where  $z_p$  is the  $p$ th quantile for standard normal distribution for  $p \leq 0.5$ . For various quantiles ( $p$ ) we will estimate various  $g_p$  and Hoaglin suggested that we take the median of all those  $g_p$  as the estimate of  $g$ . Once  $g$  is estimated, we can use this value to estimate  $h$ . For various  $p$  we can fit the following regression line

$$\log_e \frac{g(y_{1-p} - y_{.5})}{e^{-gz_p} - 1} = A + h \frac{z_p^2}{2},$$

where  $A$  is the intercept and  $h$  is the slope of the line. Thus estimate of  $h$  is the estimate of the slope of the above regression line.

### 2.2 Method of Moments

The idea of method of moments to estimate the parameters is to get as many equations as the number of parameters. Since  $gh$  family of distributions has two parameters it is enough to get two equations by setting first two sample moments equal to the first two population moments. Martinez (1984) derived the  $n$ th moment around zero of the  $gh$  distribution given by

$$E(Y^n) = \frac{1}{g^n \sqrt{1-nh}} \sum_{i=0}^n (-1)^i \binom{n}{i} e^{[(n-i)g]^2 / 2(1-nh)},$$

where,  $g \neq 0$  and  $0 \leq h \leq \frac{1}{n}$ . This gives us

$$E(Y) = \frac{1}{g\sqrt{1-h}} (e^{g^2/2(1-h)} - 1), \quad 0 \leq h \leq 1,$$

and

$$\text{Var}(Y) = \frac{1}{g^2 \sqrt{1-2h}} \left[ e^{2g^2/(1-2h)} - 2e^{g^2/2(1-2h)} + 1 \right] - \frac{1}{g^2(1-h)} \left[ e^{g^2/2(1-h)} - 1 \right]^2, \quad 0 \leq h \leq \frac{1}{2}.$$

If  $m_1$  and  $m_2$  be the first and second moments around zero of the data then we can estimate  $g$  and  $h$  by solving the following equations

$$E(Y) = m_1 \text{ and } E(Y^2) = m_2 .$$

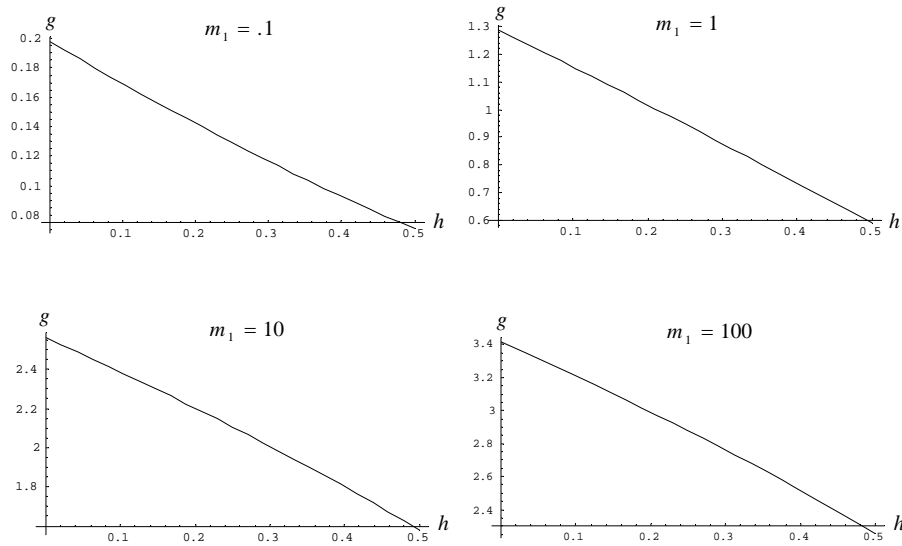
Because of the complex nature of the equations, it is quite difficult to have a closed form of the solution. Using computer system one can numerically solve the equations. But this is still a tedious job even for a computer system.

Fortunately we can estimate the solution because of a convenient nature of the equation  $E(Y) = m_1$ . Since  $m_1$  is known, this equation has two variables  $g$  and  $h$ . We can study the relationship of these two variables in this equation for various values of  $m_1$ . Figure 1 shows the plots of  $g$  for various value of  $h$ . This has been done for several values of  $m_1$  and every time it seems to be a straight line. Thus we can assume that  $g$  and  $h$  are almost linearly related. From the equation  $E(Y) = m_1$  it is possible to generate number of data pairs  $(g, h)$ . Based on this data we can have the least square estimate of  $\alpha$  and  $\beta$  where

$$g = \alpha + \beta h.$$

Then putting this value of  $g$  in the equation  $E(Y^2) = m_2$ , we can numerically solve the equation for  $h$ . Once we have the estimated value of  $h$ , putting this value in the equation  $E(Y) = m_1$  we can then numerically solve for  $g$ . One may suggest estimating  $g$  from the linear relationship obtained above for the convenience of computation. But it would be wise to estimate  $g$  from the original equation  $E(Y) = m_1$ .

Using the procedure described above we have the solution of the equations  $E(Y) = m_1$  and  $E(Y^2) = m_2$ . The results are shown in Table 2. Notice that for smaller values of the moments the solutions are very close. But for the larger values of the moments the solutions are quite close but not as close to the actual values as desirable. We can solve this problem by changing the scale of data since the variations of data does not affect the shape such as skewness and elongation. After changing the scale of data we can estimate the parameters  $g$  and  $h$  using this method and apply this to the original data.



**Figure 1:** Plot of the equation  $E(Y) = m_1$  for different values of  $m_1$

While constructing Table 2 to estimate the parameter  $\alpha$  and  $\beta$  for estimating  $g$  as a linear function of  $h$  we have considered 50 data points. The estimated values of  $m_1$  and  $m_2$  are calculated using the estimated values of  $g$  and  $h$ . Notice that they are almost close to the actual value which indicates the accuracy of this method to estimate  $g$  and  $h$ .

**Table 2:**  
Estimated Solution for  $g$  and  $h$  for Various Values of  $m_1$  and  $m_2$

Actual value		Estimated solution		Estimated	
$m_1$	$m_2$	$g$	$h$	$m_1$	$m_2$
0.01	5	0.010994	0.32893	0.01	4.999
0.1	2	0.149652	0.17206	0.1	1.998
0.1	10	0.095616	0.38707	0.1	9.996
0.1	50	0.079166	0.45985	0.1	50.310
0.5	2	1.080340	-0.23938	0.5	2.029
0.5	4	0.749211	0.08401	0.5	4.008
0.5	50	0.430426	0.39945	0.5	49.860
0.5	100	0.406047	0.42445	0.5	100.37
1.0	20	1.159870	0.09527	1.0	19.976
1.0	200	0.806966	0.35020	1.0	202.932
2.0	240	1.564020	0.10985	2.0	240.574

### 2.3 Maximum Likelihood Method

Though it is very difficult to find an explicit mathematical form of likelihood function for  $gh$  family of distributions, it is possible to find the maximum likelihood estimates numerically. In this section we present an algorithm to find the maximum likelihood estimates of  $g$  and  $h$  for the given empirical data set. Later we will simulate a random sample for specific values of  $g$  and  $h$ , and then we will apply this algorithm to obtain the maximum likelihood estimates of  $g$  and  $h$ .

Suppose the  $gh$  density is denoted by  $f_Y(y; g, h)$ . Then for a given sample  $(y_1, y_2, \dots, y_n)$  of size  $n$ , the likelihood function would be

$$L(g, h) = \prod_{i=1}^n f_Y(y_i; g, h).$$

Majumder (2007) has developed an algorithm named *CreateDensity[g,h]* described below for positive  $h$  by which we can obtain the density  $f_Y(y; g, h)$  for given  $g$  and  $h$ . Also, it is possible to obtain the density for negative  $h$  by slightly modifying this algorithm. Thus we can find the value of  $L(g, h)$  for specific  $g$  and  $h$  and a given sample data set.

#### **CreateDensity[g,h] [**

Initialize  $f_Z(z)$  as standard normal density function

**If ( $g = 0$ ) then {** Take  $Y(z) = Y_{g,h}(z) = ze^{\frac{hz^2}{2}}$  .

Derivative function  $d(z) = Y'_{g,h}(z) = e^{\frac{hz^2}{2}} + hz^2 e^{\frac{hz^2}{2}}$  }

**Else**

[Take  $Y(z) = Y_{g,h}(z) = \frac{(e^{gz} - 1)}{g} e^{\frac{hz^2}{2}}$

Derivative function  $d(z) = Y'_{g,h}(z) = e^{gz + \frac{hz^2}{2}} + hz \frac{(e^{gz} - 1)}{g} e^{\frac{hz^2}{2}}$  }

Take  $Y^{-1}(y) =$  the solution of the equation  $Y(z) = y$

Calculate Jacobian function as  $J(y) = \frac{1}{d(Y^{-1}(y))}$

Calculate  $f_Y(y) = f(Y^{-1}(y))J(y)$

Return  $f_Y(y)$  ]

The following algorithm numerically finds the value of the likelihood function. It takes  $Y, g, h$  as the input where  $Y$  is an array of size  $n$ . This array is actually the sample of size  $n$ . The algorithm is described below.

**Likelihood** $gh[g,h,Y]$  {CreateDensity $[g,h]$  Initialize  $L = 1$

Do for each sample value  $y \in Y$  {  $L = L * f_Y(y)$  } Return  $L$  }

The maximum likelihood estimates of  $g$  and  $h$  would be the values of  $g$  and  $h$  for which  $L(g,h)$  has the maximum value for a given sample data set. This can be done by plotting the likelihood function for various values of  $g$  and  $h$ .

### 3. SIMULATION OF $gh$ FAMILY OF DISTRIBUTIONS

To simulate random sample from this family we first simulate a random sample from standard normal distribution. For this we used acceptance rejection method described by Ross (2006). The algorithm that we used is described below. Note that the algorithm takes the values of  $g$  and  $h$  as input and gives a random value from the specific  $gh$  density.

**Generate** $gh(g,h) = \{$

1. Generate a random value  $U$  from  $U(0,1)$
2. Assign  $X = -\text{Log } U$
3. Generate another random value  $U_2$  from  $U(0,1)$
4. If  $U_2 \leq e^{-\frac{(X-1)^2}{2}}$  then  $Z = X$  otherwise go to step 1
5. Generate another random value  $U_3$  from  $U(0,1)$
6. If  $U_3 \leq 0.5$  then  $Z = -Z$
7. Assign  $Y = \left(\frac{e^{gZ} - 1}{g}\right) e^{hZ^2/2}$
8. Return  $Y$  }

#### 3.1 Performance of the Methods of Estimation of $g$ and $h$

Using the simulated samples we studied the performances of the parameter estimation methods. In our study of the estimation of the parameters  $g$  and  $h$ , we consider samples of size 300. We applied both method of moments and quantile method to estimate the parameters. To estimate the standard errors of the estimates 30 samples are considered.

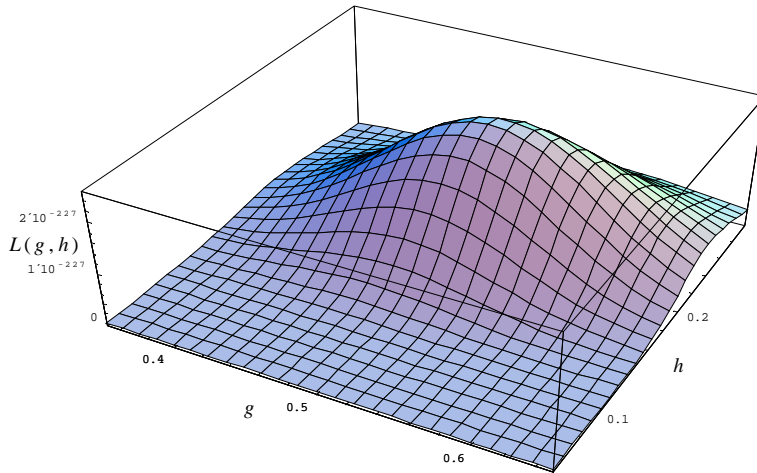
We can see from Table 3 that the estimates provided by the method of moments are much closer to the actual values of the parameters. For some cases, the estimates differ slightly from those obtained by the quantile method. This is because the solution of method of moments is not exact; instead we estimated the relationship between  $g$  and  $h$  in the first moment as linear. But still the result is acceptable since it reduces lot of computations required for the actual solution of method of moments. In addition, the estimates are as good as that of quantile method. It is evident from Table 3 that the standard errors of the estimates obtained by two methods are very close.

**Table 3:**  
**Performance of Method of Moment Algorithm**

Actual value		Method of Moments				Quantile Method			
		Estimates		Standard Error		Estimates		Standard Error	
$g$	$h$	$g$	$h$	$g$	$h$	$g$	$h$	$g$	$h$
0.5	0	0.47580	0.00319	0.13519	0.04082	0.49366	0.0145	0.11366	0.02785
0.5	0.2	0.54062	0.14817	0.15973	0.05543	0.50028	0.17867	0.12391	0.06927
0.15	0.35	0.20634	0.30200	0.20712	0.05212	0.15002	0.33757	0.09275	0.07486
0.2	0.3	0.20271	0.25758	0.15483	0.05143	0.17842	0.29107	0.14780	0.07444
0.25	0.4	0.28268	0.31274	0.22462	0.04612	0.21920	0.41506	0.12004	0.08818
0.2	-0.2	0.21733	-0.2049	0.11581	0.02702	0.19447	-0.1934	0.08759	0.02068
0.3	-0.1	0.31669	-0.1039	0.11706	0.03258	0.27419	-0.0989	0.08431	0.02181

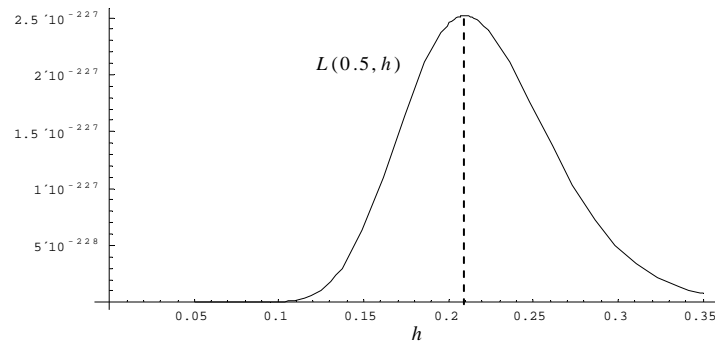
**3.2 Maximum Likelihood Estimate from Simulated Data**

To obtain the maximum likelihood estimates it requires lot of computation. Thus it is very difficult to work with large samples. In our study a sample data of size 300 is simulated from the  $gh$  density with  $g = 0.5$  and  $h = 0.2$ . To see how the algorithm for maximum likelihood estimation works, we have plotted the likelihood function as shown in Figure 4.1. Observe that there is a bump near  $g = 0.5$  and  $h = 0.2$  in Figure 2. That means the maximum likelihood estimates should be close to the true value of  $g$  and  $h$ .



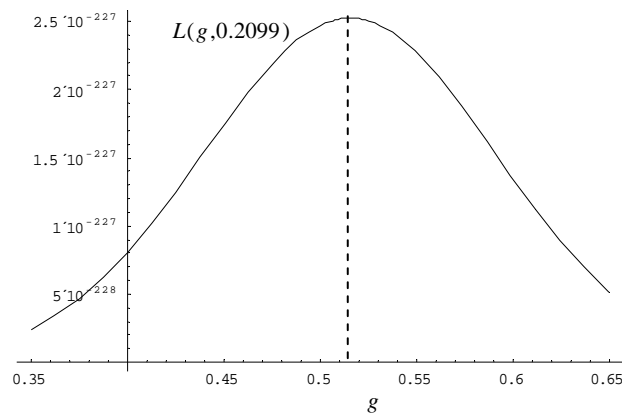
**Figure 2:** Three dimensional plot of likelihood function  $L(g, h)$

We also plotted the likelihood function  $L(g, h)$  fixing  $g = .5$  shown in Figure 3. Observe that from this plot we can estimate the value of  $h$ . For this current sample of size 300 we obtained the estimate of  $h$  as 0.2099. Figure 3 also indicates this fact.



**Figure 3:** Plot of likelihood function for  $g = .5$

We have obtained maximum likelihood estimate for  $g$  as 0.5133. Figure 4 shows the plot of likelihood function  $L(g, h)$  for  $h = 0.2099$ . Notice that the function  $L(g, h)$  for  $h = 0.2099$  is indeed maximum at point  $g = 0.5133$ . Thus we have the estimates  $g = 0.5233$  and  $h = 0.2099$ .



**Figure 4:** Plot of likelihood function for  $h = 0.2099$

#### 4. CONCLUSION

Table 4 summarizes the results obtained from our simulation process for maximum likelihood estimation. To compare this output with those in Table 3 we have considered the same actual value of  $g$  and  $h$ . Observe that the estimates are closer to the true parameters than those obtained by method of moments and quantile method. This is because maximum likelihood method gives the exact estimates while in other two methods we have to estimate the solutions. Also the standard errors are quite large for method of moments and quantile methods compared to maximum likelihood estimates.

**Table 4:**  
**Performance of Maximum Likelihood Estimate**

Actual value		MLE Method			
		Estimates		Standard Error	
$g$	$h$	$g$	$h$	$g$	$h$
0.50	0.00	0.5139	0.0019	0.0354	0.0187
0.50	0.20	0.5202	0.1823	0.0382	0.0332
0.15	0.35	0.1489	0.3453	0.0364	0.0368
0.20	0.30	0.2195	0.2921	0.0201	0.0223
0.25	0.40	0.2671	0.3730	0.0349	0.0389
0.20	-0.20	0.1862	-0.2104	0.0259	0.0256
0.30	-0.10	0.2852	-0.1123	0.0346	0.0234

#### ACKNOWLEDGEMENTS

The authors would like to thank Dr. Yulei He of the Harvard Medical School for making his Ph.D. thesis easily accessible during the conduct of the research.

#### REFERENCES

1. He, Y. and Raghunathan, T.E. (2006). Tukey's  $gh$  Distribution for Multiple Imputation, *The American Statistician*. 60(3), 251-256.
2. Hoaglin, D.C. (1985). Summarizing Shape Numerically: The  $g$ -and- $h$  distributions, in *Exploring Data Tables, Trends and Shapes*. eds. Hoaglin, D.C.; Mosteller, F. and Tukey, J.W. New York: Wiley, pp. 461-513.
3. Martinez, J. and Iglewicz, B. (1984). Some Properties of the Tukey  $g$  and  $h$  Family of Distributions, *Comm. Statist. - Theory and Methods*. 13, 353-369.
4. Majumder, M. Mahbulul, A. (2007). *On Tukey's  $gh$  Family of Distributions*. unpublished Master's Thesis, Ball State University.
5. Ross, Sheldon M. (2006). *Simulation*. 4th ed., Boston: Elsevier Academic Press, 2006.
6. Tukey, J.W. (1977). *Modern Techniques in Data Analysis*. NSF-sponsored regional research conference at Southeastern Massachusetts University, North Dartmouth, MA.