ACTUARIAL RESEARCH CLEARING HOUSE
1984 VOL. 1

Transformation of Grouped Data to Near Normality

Victor M. Guerrero and Richard A. Johnson

1. Introduction

Data may only be available in grouped form because

(i) Observations are deemed confidential as is the case with certain economic and sociological information.

(ii) Respondents to a question check the box corresponding to an appropriate interval of age, income or other quantitative variable.

The conventional way to analyze grouped observations on a single variable is to construct histograms or, sometimes, fit a normal distribution. We extend this latter approach by allowing for transformations to near normality.

Typically, transformations have been used to improve the agreement between data and the assumption of normality. Tukey (1977) gives several reasons for re-expressing data on transformed scale. His approach is to try a few selections from the transformation ladder

$$-\frac{1}{x^2}, -\frac{1}{x}, -\frac{1}{\sqrt{x}}, \log(x), \sqrt{x}, x^2$$

and then to choose the one that gives the most normal looking graph.

Box and Cox (1964) consider the family of transformations

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

which applies to non-negative observations. Using this family, which is continuous in the power λ, they present an analytical method for selecting λ.

In this paper, we investigate transformations to near normality from grouped data. When the normal approximation is adequate for the transformed data, we achieve

(i) A simple description in terms of mean and variance.

(ii) A smoothing of the data.

When the original observations are available, they should be used in the analysis. Even when there are a large number of cells, several methods of smoothing are available. Our methods will be of primary interest when there are relatively few cells and the sample size is moderate.

Out development is given in Section 2. An example appears in Section 3. Computational details are treated in Section 4 and large sample properties in Section 5. Section 6 presents an extension to life tables.

2. A Transformation of Grouped Data

The problem treated here is that of obtaining the Maximum Likelihood Estimator (MLE) when the original random variables are unobserved and the only available information is the number of observations falling within arbitrary, but specified, intervals of the real line.

Let us consider a random sample $X_1,\ldots,X_n$ from an absolutely continuous distribution with probability density function (pdf) concentrated on $(0,\infty)$.

The sample will be grouped into $k(k \geq 3)$ prespecified intervals denoted by $D_1 = [a_0,a_1)$, $D_2 = [a_1,a_2),\ldots,D_k = [a_{k-1},a_k)$ where $0 = a_0 < a_1 < \cdots < a_{k-1} < a_k = \infty$. The count of the number of observations in interval D_i will be denoted by n_i, and the total sample size is $n = \sum_{i=1}^{k} n_i$.

In order to obtain the MLE of (μ,σ,λ) we tentatively assume that there exists some λ_0 such that

$$X_j^{(\lambda_0)} = \begin{cases} \dfrac{X_j^{\lambda_0} - 1}{\lambda_0}, & \lambda_0 \neq 0 \\ \log\,(X_j), & \lambda_0 = 0 \end{cases} \qquad (2)$$

has a normal distribution. This assumption cannot hold strictly, except possibly for $\lambda_0 = 0$, because X_j is positive.

Under the normal assumption, with $\lambda \geq 0$,

$$P[X_1 \in D_i] = P[a_{i-1} \le X_1 < a_i]$$

$$= P[\frac{a_{i-1}^{\lambda}-1}{\lambda} \le \frac{X_1^{\lambda}-1}{\lambda} < \frac{a_i^{\lambda}-1}{\lambda}]$$

$$= \Phi(\frac{a_i^{(\lambda)}-\mu}{\sigma}) - \Phi(\frac{a_{i-1}^{(\lambda)}-\mu}{\sigma})$$

$$= p_i(\mu,\sigma,\lambda) \qquad (3)$$

where

$$a_i^{(\lambda)} = \begin{cases} \frac{a_i^{\lambda}-1}{\lambda}, & \text{if} \quad \lambda \neq 0 \\ \log(a_i), & \text{if} \quad \lambda = 0 \end{cases} \qquad i = 1,2,\ldots k-1$$

$$\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-u^2/2}\, du = \int_{-\infty}^{z} \phi(u)du$$

We take $p_0(\mu,\sigma,\lambda) = \Phi(\frac{a_1^{(\lambda)}-\mu}{\sigma})$ and $p_k(\mu,\sigma,\lambda) = 1 - \Phi(\frac{a_1^{(\lambda)}-\mu}{\sigma})$ for $\lambda \ge 0$. For negative λ, $p_i(\mu,\sigma,\lambda) = \Phi(\frac{a_{i-1}^{(\lambda)}-\mu}{\sigma}) - \Phi(\frac{a_i^{(\lambda)}-\mu}{\sigma})$.

The counts $(n_1,n_2,\ldots,n_k)$ then have the multinomial distribution whose probabilities are specified in (3). The log-likelihood becomes

$$\ell_n(\mu,\sigma,\lambda) = \log(n!) - \sum_{i=1}^{k} \log(n_i!) + \sum_{i=1}^{k} n_i \log[p_i(\mu,\sigma,\lambda)] \qquad (4)$$

We choose to maximize the log-likelihood using a two-stage procedure.

Step 1: Fix λ and maximize $\ell_n(\mu,\sigma,\lambda)$ with respect to μ and σ. This yields $\hat{\theta}_n(\lambda) = (\hat{\mu}_n(\lambda), \hat{\sigma}_n(\lambda),\lambda)$.

Step 2: Maximize $\ell_n(\hat{\theta}_n(\lambda))$ with respect to λ. This yields the MLE's $\hat{\theta}_n = (\hat{\mu}_n,\hat{\sigma}_n,\hat{\lambda}_n)'$.

In practice, we obtain $\hat{\mu}_n(\lambda)$, $\hat{\sigma}_n(\lambda)$ by solving

$$\frac{\partial \ell_n(\mu,\sigma,\lambda)}{\partial \mu} = -\sum_{i=1}^{k} \frac{n_i}{\sigma}\left[\frac{\phi(\frac{a_i^{(\lambda)}-\mu}{\sigma}) - \phi(\frac{a_{i-1}^{(\lambda)}-\mu}{\sigma})}{p_i(\mu,\sigma,\lambda)}\right] = 0 \tag{5}$$

$$\frac{\partial \ell_n(\mu,\sigma,\lambda)}{\partial \sigma} = -\sum_{i=1}^{k} \frac{n_i}{\sigma}\left[\frac{(\frac{a_i^{(\lambda)}-\mu}{\sigma})\phi(\frac{a_i^{(\lambda)}-\mu}{\sigma}) - (\frac{a_{i-1}^{(\lambda)}-\mu}{\sigma})\phi(\frac{a_{i-1}^{(\lambda)}-\mu}{\sigma})}{p_i(\mu,\sigma,\lambda)}\right] = 0$$

where numerator terms involving $a_0^{(\lambda)}$ and $a_k^{(\lambda)}$ are zero.

3. Application

We illustrate the technique with some data consisting of a great many observations grouped into a rather large number of cells. Our approach attempts an overall fit and it is better suited for a moderate number of observations grouped into a few cells.

We consider the life length distribution displayed in Figure 1. These data are the grouped ages of the Mexican population in 1966 which appear in Keyfitz and Flieger (1971 p. 344).

The value of λ which maximized the log-likelihood function is $\hat{\lambda} = .3500$ and for this value we obtained $\hat{\mu} = 4.775$ and $\hat{\sigma} = 2.712$. Transforming by $(a_i^{.3500} - 1)/.3500$ we draw the histogram of the transformed data with an overlay of the appropriate normal pdf in Figure 2.

The variance-covariance matrix of $(\hat{\mu},\hat{\sigma},\hat{\lambda})$ can be estimated using the results in Section 5. We find

$$\hat{\text{Var}}(\hat{\mu}) = .000000726 \qquad \hat{\text{Var}}(\hat{\sigma}) = .000000634$$

$$\hat{\text{Var}}(\hat{\lambda}) = .000000011 \qquad \hat{\text{Cov}}(\hat{\mu},\hat{\sigma}) = .000000576$$

$$\hat{\text{Cov}}(\hat{\mu},\hat{\lambda}) = .000000079 \qquad \hat{\text{Cov}}(\hat{\sigma},\hat{\lambda}) = .000000078 .$$

From Figures 1 and 2 we observe that strict normality was not achieved, although the transformation did yield a nearly symmetrical distribution. Draper and Cox (1969) noticed the fact that in some cases of ungrouped data, even when the transformation procedure does not yield normality, it helps to "regularize" data.

If we make the approximation $(X^{\lambda}-1)/\lambda$ is normal,

$$P[X \leq x] = \phi\left(\frac{\frac{x^{\lambda}-1}{\lambda} - \mu}{\sigma}\right)$$

and we approximate that X has density

$$\frac{1}{\sigma}\,\phi\left(\frac{\frac{x^{\lambda}-1}{\lambda} - \mu}{\sigma}\right) x^{\lambda-1} .$$

This approximation is graphed in Figure 1.

4. Some Computational Details

In order to find the MLE of $\underset{\sim}{\theta} = (\mu,\sigma,\lambda)'$ in our transformed normal approach, we need to use an iterative procedure and this requires some initial estimates. In practice, we can apply a specialization to our case, of a two-stage procedure proposed by Richards (1961) (the assumptions under which this procedure is valid were wrongly stated by Richards, but later Kale (1963) validated the method by assuming that the conditions of the Implicit Function Theorem hold). This method enables us to obtain the value $\underset{\sim}{\theta} = (\hat{\mu}_n,\hat{\sigma}_n,\hat{\lambda}_n)'$, for n fixed, as follows:

Step 1: Fix λ and solve the two likelihood equations

$$\frac{\partial \ell_n}{\partial \mu} = 0 , \quad \frac{\partial \ell_n}{\partial \sigma} = 0$$

This yields $\hat{\mu}_n(\lambda)$, $\hat{\sigma}_n(\lambda)$.

Step 2: Maximize the already partially maximized likelihood over λ to obtain the maximum likelihood estimates $\hat{\mu}_n$, $\hat{\sigma}_n$, $\hat{\lambda}_n$.

Once we obtained $\hat{\mu}_n(\lambda)$, $\hat{\sigma}_n(\lambda)$ for a fixed λ, we changed λ by a small amount and then $\hat{\mu}_n(\lambda)$, $\hat{\sigma}_n(\lambda)$ are used as initial values for the current stage.

Notice that $\Phi(\cdot)$ is evaluated at $(\frac{a_i^{\lambda}-1}{\lambda} - \mu)/\sigma$. For λ fixed, this is just a grouped normal data problem. Several methods have been devised for obtaining approximate MLE's for μ and σ (c.f. Lindley

(1950), Benn and Sidebottom (1976)) which may be used for getting initial values for those parameters. But the problem still remains with respect to selecting an initial value for λ .

Our calculations require the accurate evaluation of the standard normal cdf $\phi(w)$. In cases where $|w|$ is large, $|w| \geq 2.5$, we used the continued fraction approximation

$$\phi(w) = 1 - \phi(w) \left\{ \cfrac{1}{w + \cfrac{1}{w + \cfrac{2}{w + \cfrac{3}{\ddots}}}} \right\}$$

, $w > 0$

for $w > 0$ (see Abramowitz and Stegun (1964) 26.2.14 p. 932.

5. Large Sample Properties of the Estimators

Usually, when dealing with incomplete data, explicit expressions for the MLE's are not available. However, it is possible to gain some insight into the proposed procedure by studying their asymptotic properties as in Hernandez and Johnson (1979). Thus we establish strong consistency and asymptotic normality of the MLE and identify its limit with a minimum Kullback-Liebler information number property.

Let $Q = \{q_i\}_{i=1}^k$ be the true probabilities of the intervals,

$$q_i = \int_{D_i} g(x)dx \tag{6}$$

where $g(x)$ is the true pdf of X. Also, let $N = \{p_i(\mu,\sigma,\lambda)\}_{i=1}^k$ be the probabilities obtained from the approximate normal,

$$p_i = \Phi\left(\frac{\frac{a_i^\lambda - 1}{\lambda} - \mu}{\sigma}\right) - \Phi\left(\frac{\frac{a_{i-1}^\lambda - 1}{\lambda} - \mu}{\sigma}\right)$$

The Kullback-Liebler information number is defined as

$$I[Q,N] = -\sum_{i=1}^k q_i \ln \frac{p_i(\mu,\sigma,\lambda)}{q_i}$$

Notice that minus log-likelihood is of the same form but with the q_i replaced with the empirical frequencies n_i/n.

We now proceed to state and sketch the proof of the main result of this section.

Theorem 1. Let q_i be given by (6) and $p_i(\underset{\sim}{\theta})$ by (3). If

(i) the parameter space is a compact subset of $\mathbb{R}^3$

(ii) $H(\underset{\sim}{\theta}) = \sum_{i=1}^{k} q_i \log\left[\frac{p_i(\underset{\sim}{\theta})}{q_i}\right]$ has a unique global maximum as a function of $\underset{\sim}{\theta} = (\theta_1,\theta_2,\theta_3)' = (\mu,\sigma,\lambda)'$, and this is attained at $\underset{\sim}{\theta} = \underset{\sim}{\theta}_0$.

Then, $\hat{\underset{\sim}{\theta}}_n \xrightarrow{a.s.} \underset{\sim}{\theta}_0$ as $n \to \infty$.

If further,

(iii) $\underset{\sim}{\theta}_0$ is an interior point of Ω,

(iv) the Hessian of $H(\underset{\sim}{\theta})$, $\nabla^2 H(\underset{\sim}{\theta}) = \left(\frac{\partial^2 H(\underset{\sim}{\theta})}{\partial\theta_u \partial\theta_v}\right)_{3\times 3}$ is nonsingular at $\underset{\sim}{\theta}_0$.

Then, $\sqrt{n}(\hat{\underset{\sim}{\theta}}_n - \underset{\sim}{\theta}_0) \xrightarrow{d} N_3(\underset{\sim}{0}, VWV')$ as $n \to \infty$, where $V = [\nabla^2 H(\underset{\sim}{\theta}_0)]^{-1}$ and $W = (w_{uv})_{3\times 3}$ is given by (7) below.

Proof: Let $\hat{p}_{i,n} = \frac{n_i}{n}$. Stirling's approximation yields

$$\left|\frac{1}{n}\ell_n(\underset{\sim}{\theta}) - \sum_{i=1}^{k} q_i \log\left[\frac{p_i(\underset{\sim}{\theta})}{q_i}\right]\right|$$

$$\leq \left|\sum_{i=1}^{k} \hat{p}_{i,n} \log[p_i(\underset{\sim}{\theta})] - \sum_{i=1}^{k} q_i \log[p_i(\underset{\sim}{\theta})]\right|$$

$$+ \left|\sum_{i=1}^{k} \hat{p}_{i,n} \log(\hat{p}_{i,n}) - \sum_{i=1}^{k} q_i \log(q_i)\right| + o(1)$$

The right hand side converges to zero, with probability one, uniformly in $\underset{\sim}{\theta} \in \Omega$ and the consistency of $\hat{\theta}_n$ follows.

. To establish asymptotic normality, we obtain the gradient and the Hessian of the log-likelihood function by straightforward differentiation. The Hessian of $\ell_n(\underset{\sim}{\theta})$, with elements $\frac{\partial^2 \ell_n(\underset{\sim}{\theta})}{\partial\theta_u \partial\theta_v}$ for u,v=1,2,3 is readily seen to be continuous on Ω. So, using Taylor's formula,

$$\frac{1}{\sqrt{n}} \nabla\ell_n(\hat{\underset{\sim}{\theta}}_n) = \frac{1}{\sqrt{n}} \nabla\ell_n(\underset{\sim}{\theta}_0) + \frac{1}{n} \nabla^2\ell_n(\underset{\sim}{\theta}_{*n})[\sqrt{n}\ (\hat{\underset{\sim}{\theta}}_n - \underset{\sim}{\theta}_0)],$$

with $\underset{\sim}{\theta}_{*n} = \gamma_n \underset{\sim}{\theta}_0 + (1 - \gamma_n)\hat{\underset{\sim}{\theta}}_n$, $0 < \gamma_n < 1$.

We conclude that $\frac{1}{\sqrt{n}} \nabla\ell_n(\underset{\sim}{\theta}_0)$ and $-\frac{1}{n} \nabla^2\ell_n(\underset{\sim}{\theta}_{*n})[\sqrt{n}(\hat{\underset{\sim}{\theta}}_n - \underset{\sim}{\theta}_0)]$ have the same limiting distribution. Write

$$\frac{1}{n} \nabla\ell_n(\underset{\sim}{\theta}_0) = \frac{1}{n} \sum_{j=1}^{n} \left[\sum_{i=1}^{k} I_{D_i}(X_j)\underset{\sim}{\alpha}_i(\underset{\sim}{\theta}_0) \right] = \frac{1}{n} \sum_{j=1}^{n} \underset{\sim}{Z}_j(\underset{\sim}{\theta}_0)$$

where $\underset{\sim}{\alpha}_i(\underset{\sim}{\theta}_0) = (\alpha_{i1}(\underset{\sim}{\theta}_0), \alpha_{i2}(\underset{\sim}{\theta}_0), \alpha_{i3}(\underset{\sim}{\theta}_0))'$ with $\alpha_{ir}(\underset{\sim}{\theta}_0) = \left. \frac{\partial \log[p_i(\underset{\sim}{\theta})]}{\partial\theta_r} \right|_{\underset{\sim}{\theta}_0}$ for r = 1,2,3. Asymptotic normality follows from the multivariate Central Limit Theorem since the random vectors $\underset{\sim}{Z}_1(\underset{\sim}{\theta}_0), \ldots, \underset{\sim}{Z}_n(\underset{\sim}{\theta}_0)$ are iid with $E_g[\underset{\sim}{Z}_1(\underset{\sim}{\theta}_0)] = \nabla H(\underset{\sim}{\theta}_0) = \underset{\sim}{0}$ and have covariance elements

$$w_{uv} = \sum_{i=1}^{k} q_i \alpha_{iu}(\underset{\sim}{\theta}_0)\alpha_{iv}(\underset{\sim}{\theta}_0)$$

$$= \sum_{i=1}^{k} q_i \left(\left. \frac{\partial \log[p_i(\underset{\sim}{\theta})]}{\partial\theta_u} \right|_{\underset{\sim}{\theta}_0} \right) \left(\left. \frac{\partial \log[p_i(\underset{\sim}{\theta})]}{\partial\theta_v} \right|_{\underset{\sim}{\theta}_0} \right). \qquad (7)$$

□

Under the conditions of Theorem 1, $\underset{\sim}{\theta}_0$ is also that value of $\underset{\sim}{\theta}$ which minimizes the Kullback-Leibler information number $I[Q,N]$. A similar interpretation for ungrouped data is given by Hernańdez and Johnson (1981). Hence, obtaining the MLE of $\underset{\sim}{\theta}$ is asymptotically equivalent to finding the minimum of the Kullback-Leibler information number between the true probability distribution Q and a normal.

The importance of the true pdf g is reflected in the asymptotic variance of $\hat{\underset{\sim}{\theta}}_n$ derived in Theorem 1. If one is faced with the task of transforming a grouped sample to near normality, the true probabilities q_i can be estimated by the observed frequencies $\hat{p}_{i,n} = n_i/n$. Doing this, one obtains a consistent estimate of the asymptotic variance of $\hat{\underset{\sim}{\theta}}_n$ as in the application to the Mexican age distribution. For instance, the estimated (1,2) elements of V and W are

$$\hat{V}_{12} = \sum_{i=1}^{k} \frac{n_i}{n} \frac{\partial}{\partial\mu} \log p_i(\mu,\sigma,\lambda) \cdot \frac{\partial}{\partial\sigma} \log p_i(\mu,\sigma,\lambda)\Big|_{\hat{\mu},\hat{\sigma},\hat{\lambda}}$$

$$\hat{W}_{12} = \sum_{i=1}^{k} \frac{n_i}{n} \frac{\partial^2}{\partial\mu\partial\sigma} \log p_i(\mu,\sigma,\lambda)\Big|_{\hat{\mu},\hat{\sigma},\hat{\lambda}}$$

6. An Extension to Life Table Data Involving Withdrawals

In a life table setting, the basic data are

$$d_i = \text{no. deaths in } D_i = [a_{i-1}, a_i)$$

$$w_i = \text{no. withdrawals in } D_i$$

The contribution to the likelihood from the interval D_i is then

$$p_i^{d_i}(1 - p_1 - \cdots - p_{i-1})^{w_i}$$

That is, withdrawals survived at least until a_{i-1}. For the transformed data, we still have

$$p_i(\mu,\sigma,\lambda) = \Phi\left(\frac{\frac{a_i^\lambda - 1}{\lambda} - \lambda}{\sigma}\right) - \Phi\left(\frac{\frac{a_{i-1}^\lambda - 1}{\lambda} - \lambda}{\sigma}\right)$$

when $\lambda \geq 0$, so the likelihood is

$$L = \frac{n!}{\prod_{i=1}^{k} d_i! w_i!} \prod_{i=1}^{k} \left\{ p_i^{d_i}(\mu,\sigma,\lambda)\left[1 - \Phi\left(\frac{\frac{a_{i-1}^\lambda - 1}{\lambda} - \mu}{\sigma}\right)\right]^{w_i} \right\}$$

The log-likelihood is then

$$\ell_n(\mu,\sigma,\lambda) = c(\underset{\sim}{d},\underset{\sim}{w}) + \sum_{i=1}^{k} d_i \log p_i(\mu,\sigma,\lambda)$$

$$+ \sum_{L=1}^{k} w_i \log\left[1 - \Phi\left(\frac{\frac{a_{i-1}^{\lambda}-1}{\lambda} - \mu}{\sigma}\right)\right]$$

where $c(\underset{\sim}{d},\underset{\sim}{w}) = \log[n!/\sum_{i=1}^{k} d_i!\, w_i!]$. Numerically, this is only slightly more complicated to maximize than the situation where there are no withdrawals. A two-stage maximization is again applicable.

Note that we could also approximate that withdrawls occur at $(a_{i-1} + a_i)/2 = \bar{a}_i$ and then the corresponding probability would be

$$1 - \Phi\left(\frac{\bar{a}_i(\lambda) - \mu}{\sigma}\right) \quad .$$

Figure 1

Histogram and Inverted pdf in the Original Scale

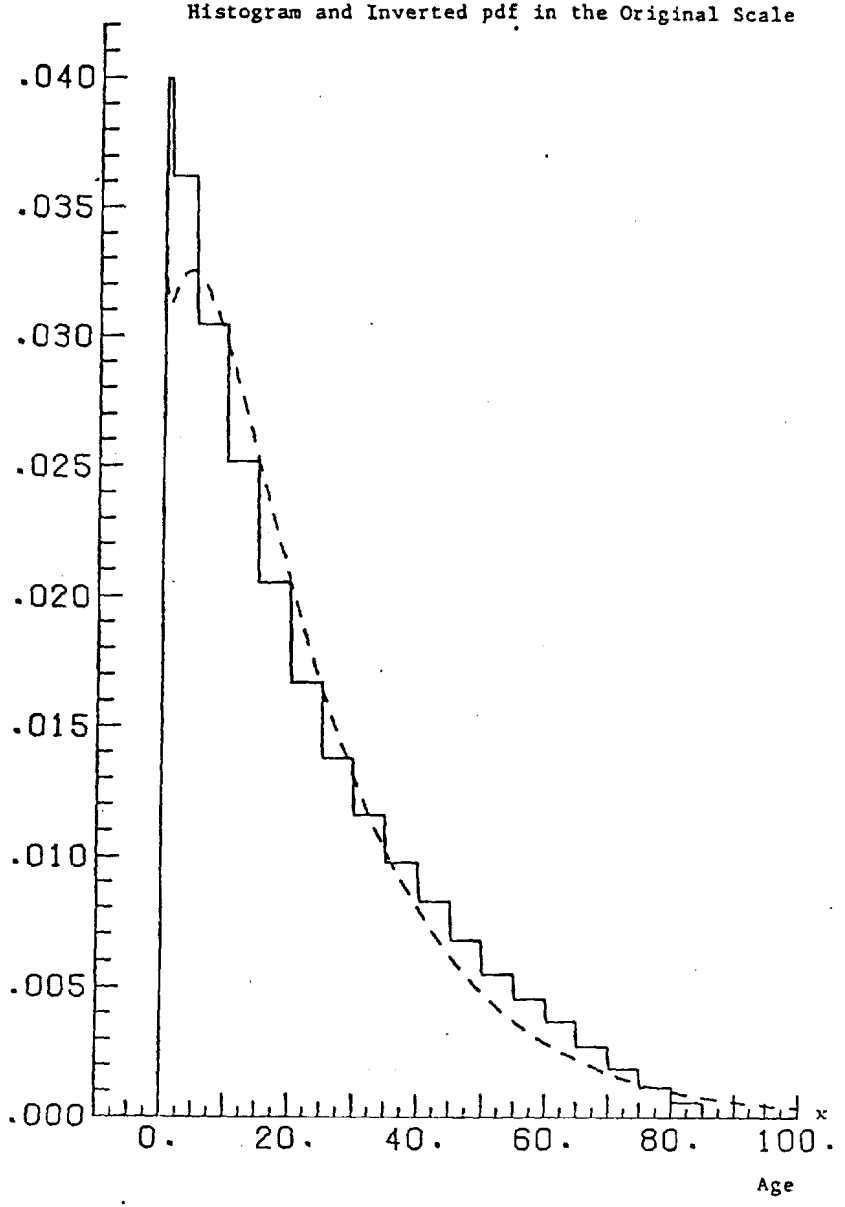
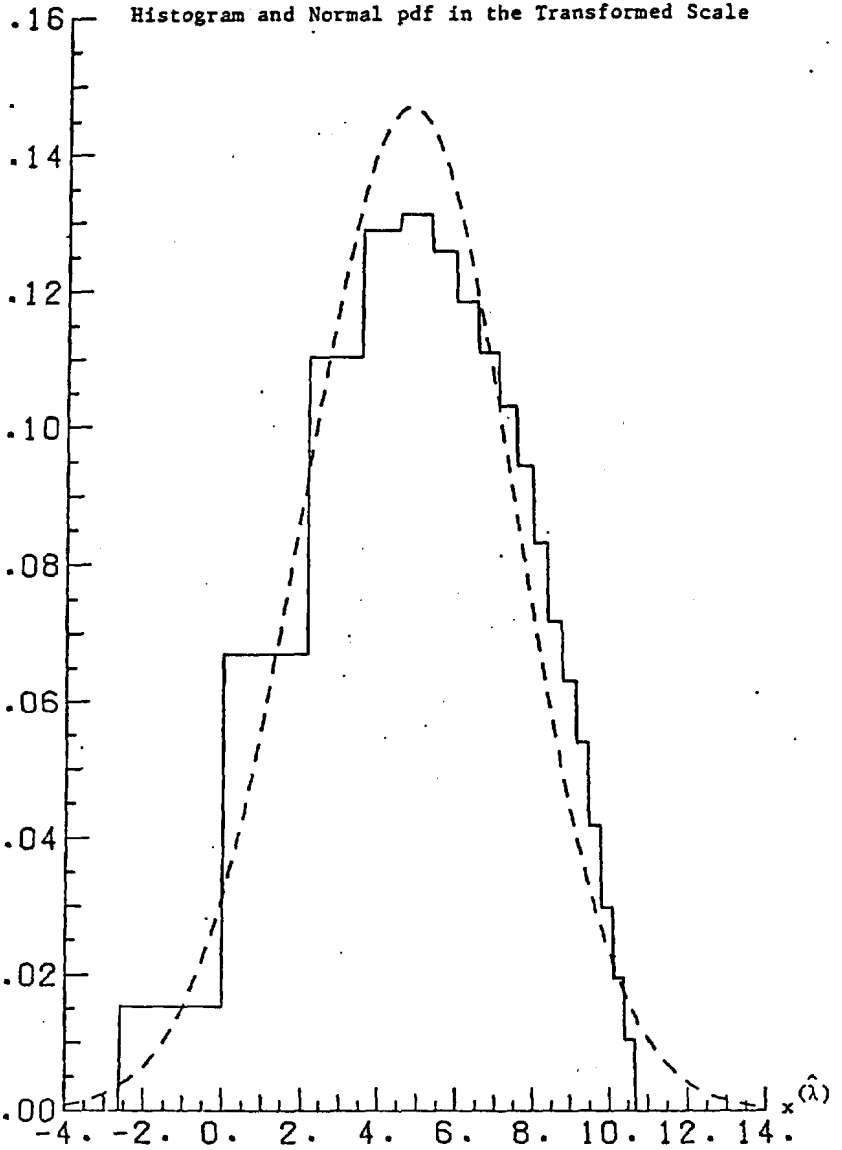


Figure 2


Histogram and Normal pdf in the Transformed Scale
.16
.14
.12
.10
.08
.06
.04
.02
.00
-4. -2. 0. 2. 4. 6. 8. 10. 12. 14.
$x^{(\hat{\lambda})}$

BIBLIOGRAPHY

Abramowitz, M. and Stegun, I. A. (1964). Handbook of Mathematical Functions, Washington, D. C.: National Bureau of Standards, Applied Mathematics Series No. 55.

Benn, R. T. and Sidebottom, S. (1976). Algorithm AS-95: Maximum likelihood estimation of location and scale parameters from grouped data, Appl. Statist. 25, 88-93.

Box, G. E. P. and Cox, D. R. (1964). An analysis of transformations, J. R. Statist. Soc. B-26, 211-52.

Draper, N. R. and Cox, D. R. (1969). On distributions and their transformation to normality, J. Royal Statist. Soc. B-31, 472-76.

Hernañdez, A. F. and Johnson, R. A. (1981). Transformation of a discrete distribution to near normality, Statistical Distributions in Scientific Work 5, 259-270, C. Taillie et. al (eds.).

Kale, K. B. (1963). Some remarks on a method of maximum-likelihood estimation proposed by Richards, J. Royal Statist. Soc. B-25, 209-12.

Keyfitz, N. and Flieger, W. (1971). Population, Facts and Methods of Demography. W. H. Freeman and Co.

Lindley, D. V. (1950). Grouping corrections and maximum likelihood equations, Proc. Camb. Philos. Soc. 46, 106-10.

Richards, F. S. G. (1961). A method of maximum likelihood estimation, J. Royal Statist. Soc. B-23, 469-75.

Tukey, J. W. (1977). Exploratory Data Analysis, Addison-Wesley, Reading, Mass.