TRANSFORMATION OF A DISCRETE DISTRIBUTION TO NEAR NORMALITY. (U)

AD-8080 866 WISCONSIN UNIV-MADISON DEPT OF STATISTICS

UNCLASSIFIED

TRANSFORMATION OF A DISCRETE DISTRIBUTION TO NEAR NORMALITY. (U)

NOV 70 P. HERNANDEZ R.A. JOHNSON

N00014-79-C-0722 NL
TRANSFORMATION OF A DISCRETE DISTRIBUTION TO NEAR NORMALITY

by

Fabian Hernández Richard A. Johnson
University of Wisconsin University of Wisconsin

DISTRIBUTION STATEMENT A
Approved for public release; Distribution Unlimited
Transformation of a Discrete Distribution
to Near Normality

by
Fabiana Hernández Richard A. Johnson
University of Wisconsin University of Wisconsin

9. Abstract

Utilizing an information number approach, we propose an objective method for the normalization of either discrete distributions, or sample counts, by means of a power transformation. Approximations are also given to the original known probabilities. Next, we derive the large sample distribution of our estimate of the power transformation. We compare our methods with the Box-Cox procedure, applied to observed counts, and conclude that their technique often provides good approximations even though their underlying assumption of normality is clearly violated. Two examples illustrate our methods.

1. Introduction and Summary

The transformation or 're-expression' of counts is now common practice for the data analyst. Tukey (1977), page 83, specifically mentions some advantages of transforming counts. Our procedure selects a 'normalizing' transformation from the family

\[
x^{(\lambda)} = \begin{cases} 
\frac{x^{\lambda+1}}{\lambda} & \lambda \neq 0 \\
\log(x) & \lambda = 0 
\end{cases} \quad x > 0
\]  

(1.1)

considered by Box and Cox. It provides an objective way to determine \( \lambda \).

In Section 2, we discuss criteria by which discrete random variables, or their transformations, may be judged to be nearly normal. Employing the Kullback-Leibler information number, we introduce a transformation technique that applies where the true underlying distribution is known. Also, we develop a discretization of the normal distribution that approximates a given discrete distribution.

Sample analogues of the methods in Section 2, are developed in Section 3 and their asymptotic distribution derived. These can be used to obtain approximate confidence intervals for the 'best' transformation parameter.

For comparative purposes, we show that the Box-Cox technique leads to sensible results, provided that we add a positive constant. Section 5 includes the re-expression of counts and the normalization of raw test scores.

2. Transformation and Approximation of a Known Discrete Distribution

2.1 Normal Approximation to transformation of Smoothed Discrete Random Variables

Let \( X \) be a discrete random variable which takes the value \( i \) with probability \( p_i = P[X = i] \) for \( i \geq 0 \). Then \( Y = X + U \) is absolutely continuous when \( U \) is independent of \( X \) and is uniform on \([c,c+1]\) some fixed \( c > 0 \).

Let \( Y \) have p.d.f. \( g(-) \) \( Y = g^{(\lambda)} \) have p.d.f. \( g^{(\lambda)}(-) \) and \( \Phi (\cdot) \) be the p.d.f. of a normal distribution with mean \( \mu \) and standard deviation \( \sigma \). In our search for a transformation, we replace the discrete variable \( X \) by the absolutely continuous \( Y \), and then select a power transformation of \( Y \).

Employing the Kullback-Leibler information number between \( g_{\lambda} \) and \( \Phi_{\mu} \), as a measure closeness, we propose to minimize

---

This research was sponsored by the Office of Naval Research under Grant No. N00014-78-C-0722.

Now at Dirección General de Estadística, Mexico.
\[ I[\theta; p_a] = E_o \left\{ \log \left( \frac{f_a(V)}{f_o(V)} \right) \right\} \]  

(2.1)

with respect to \( \mu, \sigma \) and \( \lambda \). Minimizing (2.1) first over \( \mu \) and \( \sigma \), we find that the optimal value \( \lambda_o \) of \( \lambda \) is found by minimizing

\[ G(\lambda) = \frac{1}{2} \log(2\pi) + E_o \left( \log(f_o(V)) \right) + \log\left( \frac{1}{\sqrt{2\pi}} \right) \]

(2.2)

provided that \( E_o(\lambda^2) \) and \( E_o(\log|\lambda|)^2 \) are finite. Here, \( v(\lambda) = E_o(\lambda) = E_o(\log|\lambda|)^2 \) and \( E_o(-) \) denotes that the expectation is taken with respect to \( o \).

**Proposed Procedure for Transforming a Known Discrete Distribution (2.1).**

Replace \( X \) by the absolutely continuous random variable \( Y = X + U \), where \( X \) has the given discrete distribution and \( U \), stochastically independent of \( X \), has a uniform distribution on the interval \([c, c+1)\), for some \( c > 0 \). To 'normalize' \( X \), we make the p.d.f. of \( Y \) 'closest' (in the information number sense) to a normal p.d.f. by minimizing (2.1) with respect to \( (\mu, \sigma, \lambda) \).

**Example 1** Let \( X \) have the Poisson distribution with parameter \( \alpha \). Let \( Y = X + U \) with \( U \sim U[1,2] \) (i.e. \( c=1 \)). The function \( G(\lambda) \) defined in (2.2)

\[ G(\lambda) = \text{const} - \lambda \sum_{k=0}^{\infty} \left[ \log[1 + (1/k)] + \log(2\pi) \right] p_o(1/k) - \lambda \]

\[ + \frac{1}{2} \log \left[ \sum_{k=0}^{\infty} \left( \frac{1}{k} \right)^{2\lambda+1} \left[ \log\left( \frac{1}{k} \right) \right]^{2\lambda+1} \right] p_o(1/k) \]

\[ - \left( \sum_{k=0}^{\infty} \left( \frac{1}{k} \right)^{2\lambda+1} \right) \left( \sum_{k=0}^{\infty} \left( \frac{1}{k} \right)^{2\lambda+1} - 1 \right) p_o(1/k) \left( \log(4\pi) \right) \]

for \( \lambda \neq -1 \). Here const = \[ \log(2\pi) + \sum_{k=0}^{\infty} \log(1 + (1/k)) + \log(2\pi) \] \( p_o(1/k) \) and \( p_o(1/k) \) is \( \alpha \exp(-\alpha)/(1/2) \). In Figure 1, we plot \( G(\lambda) \) vs \( \lambda \) for \( \alpha = 2, 3, 4 \) and 5.

According to our criterion, the usual variance stabilizing transformation (see Bartlett [1947], p.41), \( \lambda = \frac{1}{2} \), is a reasonable choice since it makes \( G(\lambda) \) small for these values of \( \alpha \). In Table 2.1, we record \( \lambda_o \), the information number of the transformed variable \( I[\theta_o; \rho_o] \), the information number of the untransformed variable \( I[\theta; \rho] \), \( \nu = \frac{\alpha^2}{\sigma^2} \), and \( \sigma^2 = \alpha + 1/12 \). We also include the information number \( I[\theta_o; \rho_o(h)] \) corresponding to the square root transformation.

**Table 2.1**

<table>
<thead>
<tr>
<th>( \lambda_o )</th>
<th>( I[\theta_o; \rho_o] )</th>
<th>( I[\theta; \rho] )</th>
<th>( I[\theta_o; \rho_o(h)] )</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>.36889</td>
<td>.033815</td>
<td>.08104</td>
</tr>
<tr>
<td>3</td>
<td>.40816</td>
<td>.019620</td>
<td>.05047</td>
</tr>
<tr>
<td>4</td>
<td>.54318</td>
<td>.013090</td>
<td>.03572</td>
</tr>
<tr>
<td>5</td>
<td>.57540</td>
<td>.009729</td>
<td>.02753</td>
</tr>
</tbody>
</table>

2.2 Normal Approximation of Discrete Probabilities

Alternatively we can approximate the probability \( p_i \) by a probability \( q_i \) obtained from a normal c.d.f. To measure the accuracy of this type of approximation we utilize the Kullback-Leibler information number in its discrete population version (see Kullback [1968] p. 128).
Let \( d(c,0.1) \) be fixed. Minimize the Kullback-Leibler information number

\[
I(P;Q) = P_q \log \left( \frac{P_q}{Q_q(1-q)^{-1}} \right) + \sum_{i=1}^{n} P_i \log \left( \frac{P_i}{Q_i(1-q)^{-1}} \right)
\]

(2.4)

with respect to \( Q' = (\mu, \sigma) \).

Example 2) Let \( X \) have a Binomial distribution with parameters \( n \) and \( p \) and set \( q = 1-p \). We take \( d = h \) for the normal probabilities (2.3).

In Table 2.2, we display the optimal choices \( \mu, \sigma, \lambda \) that minimize (2.4), the information number for the 'best' approximating normal probabilities, \( I(P;Q(q)) \), and the information number \( I(P;R) \) corresponding to the 'usual' normal approximation which sets \( \mu = np, \sigma = \sqrt{npq} \) and \( \lambda = 1 \). The results are given for \( p < 0.5 \) since the last two columns are symmetric about \( p = 0.5 \).

<table>
<thead>
<tr>
<th>( p )</th>
<th>( \mu_n )</th>
<th>( \sigma_n )</th>
<th>( \lambda_n )</th>
<th>( I(P;Q(q)) )</th>
<th>( I(P;R) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>0.74123</td>
<td>0.53867</td>
<td>0.56841</td>
<td>0.00003</td>
<td>0.01678</td>
</tr>
<tr>
<td>0.2</td>
<td>1.53683</td>
<td>0.75304</td>
<td>0.65421</td>
<td>0.00014</td>
<td>0.00813</td>
</tr>
<tr>
<td>0.3</td>
<td>2.37057</td>
<td>1.01300</td>
<td>0.74109</td>
<td>0.00032</td>
<td>0.00538</td>
</tr>
<tr>
<td>0.4</td>
<td>3.39250</td>
<td>1.42022</td>
<td>0.84353</td>
<td>0.00056</td>
<td>0.00238</td>
</tr>
<tr>
<td>0.5</td>
<td>4.86330</td>
<td>2.22872</td>
<td>0.97688</td>
<td>0.00087</td>
<td>0.00115</td>
</tr>
</tbody>
</table>

Notice that the procedure (2.2) fits the Binomial probabilities better when \( p \) is 'small' (or 'large'). This is opposite the performance of the 'usual' normal approximation. Table 2.3 displays \( P_i, Q_i(q_i), R \) so that comparisons may be made. Since the probabilities \( Q_i(q_i) \) and \( R \) are symmetric functions of \( p \) about \( p = 0.5 \), we display them only for \( p < 0.5 \).

3. Transformation of Counts and Normal Approximation of Observed Proportions.

3.1 Transformation of Counts.

Let \( X \) be a discrete random variable taking the value 1 with probability \( P_i = P(X_i = 1) > 0 \) for \( i = 1, \ldots, N \). Let \( X_1, \ldots, X_n \)

be i.i.d. as \( X \) and denote the frequency of the value 1 by \( f_{i1} \). Set \( I(i) = \text{the indicator of } [X_i = 1] \) so \( f_{i1} = \sum_{i=1}^{n} I(i) \). \( f_{i1} \) is the frequency of \( [X_i = 1] \) and the relative frequency is

\[
f_{i1} = f_{i1}/n \quad \text{for } i = 0, 1, \ldots, N
\]

(3.1)

Once having observed \( x_1, x_2, \ldots, x_n \), we treat these as possible values of a random variable \( Y_n \), where \( Y_n \) now takes the value 1 with probability

\[
d_{i1} = d(x_1, x_2, \ldots, x_n, Y_n = 1)
\]
Table 2.3
Comparison of Probabilities

<table>
<thead>
<tr>
<th>x</th>
<th>P = 0.1</th>
<th>P = 0.2</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.34868</td>
<td>0.34852</td>
</tr>
<tr>
<td>1</td>
<td>0.20742</td>
<td>0.39853</td>
</tr>
<tr>
<td>2</td>
<td>0.19327</td>
<td>0.195308</td>
</tr>
<tr>
<td>3</td>
<td>0.05740</td>
<td>0.05685</td>
</tr>
<tr>
<td>4</td>
<td>0.01116</td>
<td>0.01103</td>
</tr>
<tr>
<td>5</td>
<td>0.00149</td>
<td>0.00159</td>
</tr>
<tr>
<td>6</td>
<td>0.00014</td>
<td>0.00018</td>
</tr>
<tr>
<td>7</td>
<td>0.00001</td>
<td>0.00002</td>
</tr>
<tr>
<td>8</td>
<td>0.00000</td>
<td>0.00000</td>
</tr>
<tr>
<td>9</td>
<td>0.00000</td>
<td>0.00000</td>
</tr>
<tr>
<td>10</td>
<td>0.00000</td>
<td>0.00000</td>
</tr>
</tbody>
</table>

Next, construct

\[ V_n = V_{\lambda} + U_n \]

with p.d.f. \( g_{\lambda}(-) \)

where \( V_{\lambda} \) and \( U_n \) are stochastically independent and \( U_n \) has a uniform
distribution on the interval \([0, b])\), for some fixed \( b > 0\). Let \( V_n = V_{\lambda} + U_n \)
have p.d.f. \( g_{\lambda}(-) \) and \( g_{\lambda}(-) \) be the p.d.f. of a normal distribution

with mean \( \mu \) and standard deviation \( \sigma \).

Now, suppose we want to transform the data so that they appear to
come from a normal distribution. We propose to transform \( V_n \), instead of
the original observations \( x_i, i = 1, \ldots, n \) and to select a transformation
\( \lambda \) in such a way that the distribution of \( V_{\lambda} \) is 'closest' to a normal
distribution in the Kullback-Leibler information number sense. That is,
we minimize

\[ \int g_{\lambda}(x) \log g_{\lambda}(x) dx = \int g_{\lambda}(x) \log \left( \frac{g_{\lambda}(x)}{g_{\lambda}(x)} \right) dx \]

with respect to \( \mu \), \( \sigma \) and \( \lambda \).

Thus, minimizing first with respect to \( \mu \) and \( \sigma \), it is easily shown
that the optimal value, \( \lambda_{\hat{\lambda}} \), of \( \lambda \) is obtained by minimizing

\[ g_{\lambda}(x) = \frac{1}{\lambda} \log(2\pi e x) \int g_{\lambda}(x) \left( \frac{1}{\lambda} \right)^\lambda \cdot g_{\lambda}(x) \int \log(g_{\lambda}(x)) \]

[3.4]

In summary,

Proposed Procedure for Transforming Counts to Near Normality (3.1)

Having observed \( x_1, x_2, \ldots, x_n \), introduce the discrete random variable
\( Y_n \) which takes the value \( k \) with probability \( \tilde{p}_{ik} = \frac{1}{n} \int_{k-0.5}^{k+0.5} g_{\lambda}(x) dx \).
Replace $V_n$ by the absolutely continuous random variable $V_n = Y_n U_n$, where $U_n$ is independent of $V_n$, and is uniform on the interval $[0, \alpha]$. In order to 'normalize' the observations, we transform $V_n$ employing Procedure 2.1.

The transformation $\lambda$ is selected by minimizing (3.4) with respect to $\lambda$.

In order to determine the asymptotic behavior of $\bar{g}_n = (\bar{u}_{n1}, \bar{u}_{n2}, ..., \bar{u}_{nk})$, the value of $\theta$ that minimizes (3.3), we establish the following auxiliary result.

**Lemma 3.2**. Suppose $X$ is a discrete random variable with $p_j = P[X = j]$, $j = 0, 1, ..., \nu < \infty$. Let $X_1, ..., X_\nu$ be independently distributed as $X$ and $Y_n$, defined by (3.2), have p.d.f. $g_n(\cdot)$. Set $V = XU$, where $U$ is uniform on the interval $[0, \alpha]$ and independent of $X$. Let $g(\cdot)$ be the p.d.f. of $V$ and $g_n(\cdot)$ the p.d.f. of $W = V^{(1)}$. Then, with probability one

1) $\lim E_n [\log(g_n(W))]/ = E [\log(g(V))]$

2) $\lim E_n [\log(V)] = E [\log(V)]$

3) $\lim E_n [\log(\bar{g}_n(W))]/ = E [\log(\bar{g}(W))].$

**Proof**. 1) Since $E [\log(\bar{g}_n(W))] = \sum_{i=0}^{\nu} \bar{g}_n(\bar{r}_i) \log(\bar{r}_i)$, follows by (3.1) applications of the Strong Law of Large Numbers.

2) Again by the Strong Law, $E_n [\log(V)] = \sum_{i=0}^{\nu} r_i \bar{g}_n(\bar{r}_i) \rightarrow \sum_{i=0}^{\nu} r_i g_i = E [\log(V)].$

3) Also $E_n [\log(g_n(W))] = \psi(2\alpha)^2 \sum_{i=0}^{\nu} \bar{g}_n(\bar{r}_i) \log(\bar{r}_i) \rightarrow \psi(2\alpha)^2 \sum_{i=0}^{\nu} g_i(\bar{r}_i) \log(\bar{r}_i) = E [\log(\bar{g}(W))].$

**Theorem 3.3**. Let $X, Y_n, V$ and $W$ be as in Lemma 3.2. Set $\theta = (\theta_1, \theta_2, \theta_3)$ $= (\mu, \sigma, \lambda)$ and suppose that the following conditions are satisfied

1) The parameter space $\Theta$ is a compact set given by $\Theta = \{(\mu, \sigma, \lambda) : \mu \in \mathbb{R}, \sigma \in \mathbb{R}, \lambda > 0, 0 < \alpha < 1\}$.

2) $H(\theta) = I(\theta, \bar{g}, \bar{g}_n)$ has a unique minimum at $\bar{\theta}_n = (\bar{\theta}_{n_1}, \bar{\theta}_{n_2}, \bar{\theta}_{n_3})^{(0)}$.

Then,

1) $\lim \bar{\theta}_n = \bar{\theta}$, with probability one.

If we further assume that

3) $\bar{\theta}$ is an interior point of $\Theta$.

4) $V^2 H(\bar{\theta}) = \left(\begin{array}{c}
-\frac{2}{\psi(2\alpha)^2} H(\bar{\theta})^{-1}
\end{array}\right)$ is non-singular.

Then

2) $\sqrt{\psi(2\alpha)^2} \bar{\theta}_n \rightarrow N_3(0, \Sigma)$ where $V = [V^2 H(\bar{\theta})]^{-1}$ and $\Sigma$ is defined in (3.5).

**Proof**. According to (3.3) and Lemma 3.2

$E_n [(\theta_n - \theta_0) \cdot (\theta_n - \theta_0)] = E_n [\log(g_n(W))] + (1-\lambda) E_n [\log(V)] - E_n [\log(\bar{g}_n(W^{(1)}))]$

$= \psi(2\alpha)^2 + E_n [\log(g_n(W))] + (1-\lambda) E_n [\log(V)] + E_n [\log(\bar{g}(W))] - E_n (\log(\bar{g}_n(W^{(1)}))].$
with \( c_1(1,0) = \frac{1}{10} \frac{3}{10} \frac{1}{10} \frac{1}{10} \) and \( c_1(1,0) = \frac{1}{10} \frac{1}{10} \frac{1}{10} \frac{1}{10} \).

Consequently,

\[
\phi_n(1,0) \sim N(0, \sigma^2),
\]

and since the \( (t,s) \)

entry of \( \phi_n(1,0) \) is of the form

\[
\frac{1}{\sigma^2} b_0^T \phi_n(1,0) b_0,
\]

uniformly continuous on \( N \),

the \( \phi_n(1,0) \) is nonsingular, and setting

\[
\phi_n(1,0) - \frac{N(0, \sigma^2)}{\sigma^2} b_0^T \phi_n(1,0) b_0,
\]

the result follows from Slutsky's theorem. 

Moreover, \( \phi_n(1,0) \) is nonsingular, and setting

\[
\phi_n(0,0) = \frac{1}{\sigma^2} b_0^T \phi_n(1,0) b_0,
\]

the result follows from Slutsky's theorem. 


e. Theorem 3.1 says that \( \phi_n \) converges, with probability one, to \( \phi \) the value of \( \phi \) that minimizes the kullback-leibler information number between \( \phi_n \) the r.d.f. of \( \hat{f}(n) \), and a normal p.d.f. Hence, procedure (1.1) can be interpreted as the infinite-sample analogue of the current technique. 

3.2 A Normal Approximation to Observations Proportions

We want to approximate observed proportions (3.1) by a set of normal

numbers \( q(n) = (n, n, \ldots, n) \), \( 0 \leq q \leq n \), given by (3.2).

\[
q(n, n, \ldots, n) = \left( \begin{array}{c}
\sum_{i=1}^{n} x_i \\
\sum_{i=1}^{n} y_i \\
\end{array} \right)
\]

where \( x_i \) and \( y_i \) are independent.

Hence, \( \hat{f}(n) \) and \( \hat{f}(n) \) have the same limiting distribution. The multivariate central limit theorem applied to \( \hat{f}(n) \) yields

\[
\hat{f}(n) \sim N(0, \sigma^2),
\]

where

\[
\hat{f}(n) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} x_i
\]

and

\[
\hat{f}(n) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} y_i
\]

with

\[
\hat{f}(n) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (x_i - n)
\]

and

\[
\hat{f}(n) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (y_i - n)
\]

for the last terms. But
Proposed Procedure to Approximate Observed Proportions (3.4) In order to approximate $\hat{p}_i = \{\hat{p}_i ; \Omega \in \Omega \}$ by a collection of normal probabilities (2.3), we minimize the Kullback-Leibler information number

$$I(\hat{p}_i; \Omega(0)) = \frac{N}{1} \sum_{i=0}^{q} p_i \log \frac{p_i}{q_i(0)}$$

with respect to $\Omega$. \(\Box\)

Theorem 3.5 Let $X$ assume the value 1 with probability $p_i$ for $i = 0, 1, \ldots, N$ where $N$ is finite. Let $X_1, \ldots, X_n$ be i.i.d. with the same distribution as $X$ and $\hat{p}_i$ be the observed proportion of the value $i$. Set $\theta^* = (\theta_2, \theta_3, \theta_2) = (\mu, \sigma, \lambda)$ and assume that the following conditions are satisfied.

1) The parameter space $\theta$ is as in Theorem 3.3.

2) $F(\theta) = I(\hat{p}_i; \Omega(0)) = \frac{N}{1} \sum_{i=0}^{q} p_i \log \frac{p_i}{q_i(0)}$ has a unique minimum at $\theta_0$.

Then, $\hat{\theta}_n$, the value which minimizes (3.6), satisfies

1) $\lim_{n \to \infty} \hat{\theta}_n = \theta_0$, with probability one.

If we further assume that

2) $\hat{\theta}_n$ is an interior point of $\theta$
3) $\nabla^2 F(\theta_0) = \left[ \frac{\partial^2 F(\theta_0)}{\partial^2 \theta_j} \right]_{j, k=0}$ is non-singular.

Then

2) $\nabla^2 F(\theta_0) \to_N \theta_3(0, \theta_0)$ where $V = \left[ \nabla^2 F(\theta_0) \right]^{-1}$ and $W$ is defined in (3.7).

Proof

1) According to (3.6)

$$I(\hat{p}_i; \Omega(0)) = \frac{N}{1} \sum_{i=0}^{q} p_i \log \frac{p_i}{q_i(0)} \to_N I(\hat{p}_i; \Omega(0))$$

with probability one. But $q_i(0)$ is uniformly continuous in $\Omega(\theta)$, so the convergence is uniform in $\theta$. Consequently, $\lim_{n \to \infty} \hat{\theta}_n = \theta_0$, with probability one.

2) To establish the limiting normality of $\hat{\theta}_n$, we introduce the function

$$F_n(\theta) = I(\hat{p}_i; \Omega(0))$$. A Taylor expansion of $\nabla F_n(\theta_0)$ yields

$$\nabla F_n(\theta_0) = \nabla F_n(\theta_0) + \nabla^2 F_n(\theta_0) \left[ \theta_0 - \theta_0 \right]$$

where $\theta_0 = \left[ \hat{\theta}_n + (1 - \hat{\theta}_n) \hat{\theta}_n \right]$ with $\hat{\theta}_n \in (0,1)$. Next, since $\hat{\theta}_n \to \theta_0$ for large enough $n$. Hence, $\nabla F_n(\theta_0)$ and $\nabla^2 F_n(\theta_0) \left[ \theta_0 - \theta_0 \right]$ have the same asymptotic distribution. Moreover, the Multivariate Central Limit Theorem applied to $V F_n(\theta_0)$ yields

$$\nabla F_n(\theta_0) \to_N \theta_3(0, \theta_0)$$

where

$$W = \left[ \frac{N}{1} \sum_{i=0}^{q} \frac{\log q_i(0)}{\partial^2 \theta_j} \right]_{j, k=0}$$

Consequently, $\nabla^2 F_n(\theta_0) \left[ \theta_0 - \theta_0 \right] \to_N \theta_3(0, \theta_0)$. It is easily shown that $\nabla^2 F_n(\theta) \to_N \theta_3(0, \theta_0)$ uniformly, with probability one.

Setting $V = \left[ \nabla^2 F_n(\theta_0) \right]^{-1}$ and utilizing Slutsky's theorem

$$V(\nabla^2 F_n(\theta_0) \left[ \theta_0 - \theta_0 \right]) \to_N \theta_3(0, \theta_0)$$

and hence,

$$\nabla^2 F_n(\theta_0) \to_N \theta_3(0, \theta_0) \Box$$
4. A Comparison with the Box-Cox Procedure Applied to Discrete Observations.

In our model \( X = 1 + C \) with \( c > 0.5 \), the Box and Cox (1964) method selects \( \lambda \) by maximizing

\[
\ell_{\text{max}}(\lambda) = - \frac{N}{2} [\log(n)+1] - \frac{N}{2} \log \left( \frac{1}{n} \sum y_i^{(1)} \right)^2 - (n-1) \log (\lambda)
\]

with respect to \( \lambda \).

**Lemma 4.1** Let \( \theta \) be the parameter space be restricted as in Theorem 3.3. Then the MLE \( \hat{\lambda}_n \), obtained by maximizing (4.1) converges with probability one to \( \lambda_0 \), the value of \( \lambda \) that maximizes

\[
g(\lambda) = \lambda \{ \log(\lambda) - \log(\lambda)^2 \} - \{ \log(\lambda)^2 \} + \log(\lambda)
\]

*Proof* Since \( \theta \) is compact and \( Y \) only assumes a finite number of values, it follows from Rubin (1956) that the probability one convergence of

\[
\hat{\lambda}_n \rightarrow \lambda_0
\]

is uniform in \( \lambda \). Consequently \( \hat{\lambda}_n \rightarrow \lambda_0 \). \( \square \)

Let \( Y = X + U \) where \( U \) is independent of \( X \) and is uniform on \( [c-h, c+h] \). Procedure 2.1, the large sample limit of Procedure 3.1, selects \( \lambda \) to minimize \( g(\lambda) \). Alternatively, the Box-Cox approach, applied directly to the discrete observations, requires the minimization of \( g(\lambda) \).

It can be shown that

\[
e(\lambda) = -g(\lambda) + \text{constant} + \text{Error}
\]

where

\[
\text{Error} = \lambda \frac{1}{n} \sum_{i=0}^{n-1} \frac{1}{y_i^{(1+c)}} \left( \log(1+c) - \log(y_i) \right)
\]

\[
- \frac{1}{4} \log \left( \frac{1}{n} \sum_{i=0}^{n-1} \frac{1}{y_i^{(1+c)}} \left( \log(1+c) - \log(y_i) \right)^2 \right)
\]

\[
\frac{1}{4} \left( \frac{1}{n} \sum_{i=0}^{n-1} \frac{1}{y_i^{(1+c)}} \left( \log(1+c) - \log(y_i) \right)^2 \right)^2
\]

\[
\frac{1}{4} \left( \frac{1}{n} \sum_{i=0}^{n-1} \frac{1}{y_i^{(1+c)}} \left( \log(1+c) - \log(y_i) \right)^2 \right)^2
\]

When

\[
\left| \int_{1+c-h}^{1+c} \left( \log(y) - \log(1+c) \right) dy \right| \leq \frac{1}{1+c-h}
\]

and, for \( r=1,2 \)

\[
\left| \int_{1+c-h}^{1+c} \left( \left( 1+c \right)^{y_i} \left( 1+c \right)^{2r} \right) dy \right| \leq \frac{1}{2} \left( \left( 1+c-h \right)^{y_i} \right)
\]

are small, we would expect the Error to be small. Consequently the Box-Cox procedure and Procedure 3.1 should give nearly the same answer when the sample size is large.

From (4.4) and (4.5) making \( c \) large appears to improve the agreement between the two procedures. The most frequently employed transformations \( \lambda \), \( \lambda^{1/2}, \log(1+c) \) and \( (1+c)^{-1} \) satisfy \( x \lambda \).

5. Application to the Re-expression of Counts and Normalization of Test Scores

**Example 1** Tukey (1977), page 572, displays the following counts for the duration (in days) of incubation for 1663 eggs of the ridley turtle.
We suggest the re-expression of the data. Setting \( \beta = 0 \), we apply the Proposed Procedure 3.1 to the above data and obtain \( \hat{\theta}_{1663} = (3.570, 1.333, 0.015)^T \). Moreover, using the limiting distribution of \( \hat{\theta}_{0} \), derived in Theorem 3.3, we can establish an approximate confidence interval for \( \theta_{0} \). The estimated standard error of \( \hat{\theta}_{1663} \) is \( \sqrt{0.015} \). Hence, an approximate 95% confidence interval for \( \theta_{0} \) is \( (0.597, 0.933) \). Notice that \( \theta_{0} \) is not included in the interval. □

**Example 2** Ghiselli (1964), page 78, proposes the use of the square root transformation for the normalization of the 100 test scores.

We set \( \beta = 0 \). The application of the Proposed Procedure 3.1 yields \( \hat{\theta}_{100} = (4.46, 0.54, 0.070) \). Utilizing the limiting distribution of \( \hat{\theta}_{0} \), we obtain the estimated standard error of \( \hat{\theta}_{100} \) and then the approximate 95% confidence interval \( (-0.30, 0.44) \) for \( \hat{\theta}_{0} \). Figure 3 presents a comparison of the relative frequency histograms of a) the transformed scores and b) the original scores. We also applied the Box-Cox procedure to the above scores. The estimated power transformation is \( \lambda = 0.078 \) which is in good agreement with the value \( \lambda_{100} = 0.070 \). □

The sample procedures introduced in Section 3, can also be applied to situations where the observations can only be ordered. That is, the observations can be assigned to exactly one of the categories 1, 2, ..., \( M \).

Here \( M \) corresponds to the highest category, \( M-1 \) to the second highest, etc. With this scoring, we are able to apply our methods to the relative frequencies of the categories.
REFERENCES


**Title:** Transformation of a discrete distribution to near normality

**Authors:** Fabian Hernandez, Richard A. Johnson

**Abstract:** Utilizing an information number approach, we propose an objective method for the normalization of either discrete distributions, or sample counts, by means of a power transformation. Approximations are also given to the original known probabilities. Next, we derive the large sample distribution of our estimate of the power transformation. We compare our methods with the Box-Cox procedure, applied to observed counts, and conclude that their technique often provides good approximations even though their underlying assumption of normality is clearly violated. Two examples illustrate our methods.