TRANSFORMATION THEORY:
HOW NORMAL IS A FAMILY OF DISTRIBUTIONS?

BY

BRADLEY EFRON

TECHNICAL REPORT NO. 173
AUGUST 1981

PREPARED UNDER THE AUSPICES
OF
NATIONAL SCIENCE FOUNDATION GRANT MCS80-24649

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
TRANSFORMATION THEORY:
HOW NORMAL IS A FAMILY OF DISTRIBUTIONS?

Bradley Efron
Stanford University

ABSTRACT
This paper concerns the following question: $X$ is a real-valued random variate having a one-parameter family of distributions $\mathcal{F}$. To what extent can $\mathcal{F}$ be normalized by a monotone transformation? In other words, does there exist a single transformation $Y = g(X)$ such that $Y$ has, nearly, a normal distribution for every distribution of $X$ in $\mathcal{F}$. The theory developed answers the question without considering the form of $g$ at all. In those cases where the answer is yes, simple formulas for calculating $g$ are given. The paper also considers the relationship between normalization and variance stabilization.
TRANSFORMATION THEORY:
HOW NORMAL IS A FAMILY OF DISTRIBUTIONS?

Bradley Efron
Stanford University

1. **Introduction.** The classic example of a normalizing transformation concerns the correlation coefficient. If \( \theta \) is the correlation of a bivariate normal distribution, and \( X \) the sample correlation of \( n \) independently drawn points from this distribution, then

\[
Y = \tanh^{-1} X = \frac{1}{2} \log \frac{1+X}{1-X}
\]

has, approximately, a normal distribution

\[
Y \sim N(\nu_\theta, \frac{1}{n-3})
\]

where

\[
\nu_\theta = \tanh^{-1} \theta + \frac{\theta}{2(n-1)}.
\]

Hotelling (1953) extensively discusses approximations (1.2), (1.3), and their higher-order improvements. Transformation (1.1) was originally suggested by Fisher (1915).

Why was Fisher interested in transforming the family of correlation distributions? (1) Quick calculations of significance levels are much easier on the \( Y \) scale. For example, with \( n = 15 \) is \( x = 0.70 \) significantly different from \( \theta = 0.20 \)? Yes: \( \sqrt{n-3} (y-\nu_{0.2}) = 2.277 \), significant at level \( .011 \), one-sided. (2) Normal theory methods can be applied on the \( Y \) scale. Suppose \( X_1 = x_1, X_2 = x_2, \ldots, X_j = x_j \) are observed sample
correlations calculated from independent data sets, and we are interested in the relationship of these correlations to covariate vectors \( c_1, c_2, c_3, \ldots, c_J \). A standard linear regression analysis of the transformed values \( y_1, y_2, \ldots, y_J \) versus \( c_1, c_2, \ldots, c_J \) is the natural way to proceed. (3) Confidence intervals are obvious on the \( Y \) scale. The 90% central interval \( v_\theta \in y \pm 1.645/\sqrt{n-2} \) gives the usual interval for \( \theta \), by inverting function (1.3).

In fact, there is a more basic reason for transformation theory's hold upon the interest of statisticians. The fundamental mathematical unit of statistical inference is a family of probability distributions. Fisher's transformation relates a complicated-looking family \( \mathcal{J} \), the correlation distributions, to a simpler family \( \mathcal{Q} \), the normal translation family. The appeal of (1.1) - (1.3) is similar to representing a symmetric matrix \( F \) as \( \Gamma G \Gamma' \), where \( \Gamma \) is orthogonal and \( G \) is diagonal. We feel, correctly, that we have increased our understanding of \( F \) by the representation in terms of \( G \), and likewise in the case of \( \mathcal{J} \) and \( \mathcal{Q} \).

This paper concerns the following question: \( X \) is a real-valued random variate having a one-parameter family of distributions \( \mathcal{J} \). To what degree can \( \mathcal{J} \) be normalized? In other words, does there exist a single monotone transformation \( Y = g(X) \) such that \( Y \) has, nearly, a normal distribution for every distribution of \( X \) in \( \mathcal{J} \)?

It seems as if we have to examine all possible monotone transformations \( Y = g(X) \) in order to answer the question. In fact it is not necessary to consider \( g \) at all. If a normalizing \( g \) exists then the cumulative distribution function \( F_\theta(x) \) of \( X \) must be of the form
\[ F_\theta(x) = \Phi \left( \frac{g(x) - \nu_\theta}{\sigma_\theta} \right). \] (1.4)

Here \( \Phi \) is the standard normal cdf, \( \theta \) is the real-valued parameter indexing \( \mathcal{F} \), \( \nu_\theta \) is the median of \( g(X) \), and \( \sigma_\theta \) the standard deviation.

For two different values of \( \theta \), say \( \theta_1 \) and \( \theta_2 \), define
\[ z_i(x) = \Phi^{-1} \left( F_{\theta_i}(x) \right), \] (1.5)
i = 1, 2, so that by (1.4)
\[ z_i(x) = \frac{g(x) - \nu_{\theta_i}}{\sigma_{\theta_i}}. \] (1.6)

Eliminating \( g(x) \) from the two equations (1.6) gives
\[ z_2(x) = \frac{\sigma_{\theta_2}}{\sigma_{\theta_1}} z_1(x) + \frac{\nu_{\theta_1} - \nu_{\theta_2}}{\sigma_\theta}. \] (1.7)

The quantities \( z_i(x) \) can be calculated directly from the cdf's \( F_{\theta_i}(x) \), without any knowledge of \( g \). Equation (1.7) shows that if \( \mathcal{F} \) can be normalized, then \( z_2(x) \) is a linear function of \( z_1(x) \).

Figure 1 plots \( z_2(x) \) versus \( z_1(x) \) for the normal correlation family originally considered by Fisher, \( n = 15, \theta_1 = 0.5, \theta_2 = 0.7 \). The plot is nearly, but not perfectly, linear. Moreover the slope of the fitted straight line is nearly 1. From (1.7) we see that this implies \( \sigma_{\theta_2}/\sigma_{\theta_1} \approx 1 \), in agreement with (1.2).

The diagnostic function \( D(z, \theta) \) introduced in Section 2 is a more convenient way of carrying out the same calculation, without having to consider all pairs of values \( \theta_1 \) and \( \theta_2 \). It enables us to diagnose deviations from the ideal of perfect normalization. For example, the normal correlation family, \( n = 15 \), is better represented as a monotone
transformation of a translation family in which the basic distribution is Student-t with 38 degrees of freedom, rather than perfectly normal (Section 4).

Figure 1. A plot of $z_2(x)$ versus $z_1(x)$ for the normal correlation family, $n = 15$, $\theta_1 = 0.5$, $\theta_2 = 0.7$. The plot is nearly linear, indicating the existence of a nearly normalizing transformation.
Fisher's transformation for the correlation coefficient is stabilizing as well as normalizing. There is no function \( \sigma_\theta \) in (1.2). The \( \tanh^{-1} \) transformation produces stable variances as well as normality. For other families \( \mathcal{F} \) there is tension between the twin goals of stabilization and normalization. In the Poisson family, \( \sqrt{x + \frac{3}{8}} \) is an excellent variance stabilizer, while \( x^{2/3} \) is an excellent normalizer, Anscombe (1948, 1953).

The second purpose of this paper is to examine the relationship between stabilization and normalization. For example, we show that the ideal stabilizing transformation for the Poisson family goes about 37% past the ideal normalizer, in a sense made precise in Section 7. In order to stabilize the Poisson family we have to transform past normality. These calculations are related to those in Tukey (1958). Simple formulas for the normalizing and stabilizing transformations, for any family \( \mathcal{F} \), are given in Section 5. Section 8 concludes the paper with a brief discussion of the relative merits of stability versus normality.

2. A Diagnostic Function. We are given \( \mathcal{F} \), a one-parameter family of distributions for the real-valued continuous variate \( X \). Let

\[
F_\theta(x) = \text{Prob}_\theta \{ X \leq x \}
\]

be the cumulative distribution function of \( X \) for parameter value \( \theta \), where \( \theta \in \Theta \) the parameter space, a possibly infinite interval of the real line. The derivatives \( F'_\theta(x) = \frac{\partial}{\partial \theta} F_\theta(x) \) and \( f_\theta(x) = \frac{\partial}{\partial x} F_\theta(x) \) are assumed to exist in what follows. We wonder whether \( \mathcal{F} \) is a normal transformation family, abbreviated Ntf, that is whether there exists a strictly monotonic transformation \( g(x) \) such that
\[ g(X) \sim h(\nu_\theta, 1) \]  

(2.2)

for all \( \theta \in \Theta \). Here \( \nu_\theta \) is the center of the normal distribution for \( g(X) \) under parameter value \( \theta \).

To answer this question, we construct a diagnostic function \( D(z, \theta) \) in the following way: let \( \hat{F}_\theta(x) = \frac{\partial}{\partial \theta} F_\theta(x) \), and, for \( 0 < \alpha < 1 \), define

\[ x_{\alpha, \theta} : F_\theta(x_{\alpha, \theta}) = \alpha, \]  

(2.3)

so that \( x_{\alpha, \theta} \) is the 100th percentile point for \( X \) under \( F_\theta \). In particular, \( x_{0.5, \theta} \) is the median of \( X \). Then the diagnostic function is defined as

\[ D(z, \theta) = \frac{\hat{F}_\theta(x_{\Phi(z), \theta}) \phi(\Phi(z))}{\hat{F}_\theta(x_{0.5, \theta}) \phi(z)}, \]  

(2.4)

with \( \phi(z) = (2\pi)^{-1/2} \exp(-z^2/2) \) and \( \Phi(z) = \int_{-\infty}^{z} \phi(z')dz' \) as usual.

Notice that \( D(z, \theta) \) is defined directly in terms of the cdf's \( F_\theta(x) \), so that it can be evaluated without knowledge of \( g(x) \), or even the assumption that \( g(x) \) exists. Some motivation for definition (2.4) is given near the end of this section.

It turns out that \( D(z, \theta) = 1 \) if \( \Theta \) is a normal transformation family. More usefully, plots of \( D(z, \theta) \) enable us to diagnose deviations of \( \Theta \) from the ideal form (2.2). To this end consider a more general family \( \Theta \) satisfying

\[ g(X) \sim \nu_\theta + \sigma_\theta q(Z) \]  

(2.5)

for some strictly monotonic transformation \( g(x) \). Here \( Z \sim h(0,1) \) is a standard normal deviate; \( q(z) \) is a strictly increasing differentiable function satisfying.
\[ q(0) = 0 ; \quad q'(0) = 1 ; \quad (2.6) \]

and \( v_{\theta} \) and \( \sigma_{\theta} > 0 \) are differentiable functions of \( \theta \), not necessarily monotonic, though we assume \( \dot{v}_{\theta} = \partial v / \partial \theta \neq 0 \) except at a finite number of \( \theta \) values.

For a normal transformation family (2.2), \( \sigma_{\theta} = 1 \) and \( q(z) = z \). The form (2.5) allows the scaling parameter \( \sigma_{\theta} \) to vary with \( \theta \), and also for \( g(X) \) to be a location-scale transformation of a general variate \( \tilde{Z} = q(Z) \) rather than just a normal variate \( Z \). We call a family \( \mathcal{F} \) satisfying (2.5) for some choice of \( g(x) \), \( q(z) \), \( v_{\theta} \), and \( \sigma_{\theta} \) a general scaled transformation family, abbreviated \( G'\text{tf} \).

**Lemma 1.** If \( \mathcal{F} \) is a general scaled transformation family, then the diagnostic function equals

\[ D(z, \theta) = \frac{1 + q(z) \varepsilon_{\theta}}{q'(z)} , \quad (2.7) \]

where

\[ \varepsilon_{\theta} = \frac{\dot{\sigma}_{\theta}}{v_{\theta}} = \frac{\partial \sigma_{\theta}}{\partial \theta} \frac{1}{\partial v_{\theta} / \partial \theta} . \quad (2.8) \]

(Proof later in this section.)

If \( \mathcal{F} \) is a normal transformation family then \( \sigma_{\theta} = 1 \), \( \varepsilon_{\theta} = 0 \), \( q(z) = z \), \( q'(z) = 1 \), and so \( D(z, \theta) = 1 \) as claimed.

Suppose \( q(z) = z \) but that \( \sigma_{\theta} \) is not a constant, a situation called a normal scaled transformation family, \( N'\text{tf} \), in Section 3.

Then the lemma gives

\[ D(z, \theta) = 1 + z \varepsilon_{\theta} . \quad (2.9) \]
As an example consider the continuous Poisson family

\[ \int_0^\infty t^{x-1/2} e^{-t} dt \frac{\theta}{\Gamma(x + \frac{1}{2})}, \quad x > -\frac{1}{2}, \tag{2.10} \]

\( \theta = (0, \infty). \) If \( x_0 \) is a nonnegative integer then

\[ F_\theta(x_0 + \frac{1}{2}) = \text{Prob}[G_{x_0+1} > \theta] = \text{Prob}[\text{Po}(\theta) < x_0], \tag{2.11} \]

where \( G_{x_0+1} \) is a gamma variate with shape parameter \( x_0+1 \) and \( \text{Po}(\theta) \) is a standard Poisson variate with parameter \( \theta. \) (The last equality in (2.11) is well-known from the theory of Poisson processes.) In other words, the cdf of a continuous Poisson with parameter \( \theta \) agrees with the cdf of a standard Poisson \((\theta)\) at every half-integer point. Our transformation theory applies to continuous variates, but we argue in Section 7 that the main implications apply to the standard Poisson family. Blom (1954) uses a similar device.

Figure 2 shows \( D(z, \theta) \) for the continuous Poisson family. Good agreement with (2.9) is evident, even for small values of \( \theta. \) The function \( e_\theta = c_0/\nu_\theta \) declines from .15 at \( \theta = 1 \) to .04 at \( \theta = 15. \) In other words, there is a monotonic mapping which nearly normalizes the continuous Poisson, but the normalized family has standard deviation \( \sigma_\theta \) increasing rather rapidly as a function of the median \( \nu_\theta, \) especially for small values of \( \theta. \) These points are discussed further in Sections 4 and 6. Section 7 gives a more complete discussion of the Poisson family.

The lemma enables us to calculate \( q(z) \) from \( D(z, \theta) \) after which we can also calculate the functions \( g(x), \nu_\theta, \) and \( \sigma_\theta, \) see Sections 4, 5, and 6. This means that the G'lf representation (2.5) of a family \( \mathcal{F} \) is unique, with one interesting exception discussed in Section 4. The
Figure 2. The diagnostic function $D(z, \theta)$ for the continuous Poisson family. Insert shows $\varepsilon_\theta = \bar{\sigma}_\theta / \bar{\nu}_\theta$.

restrictions (2.6) on $q(z)$ are necessary to avoid trivial nonuniqueness.

Suppose we make a monotonic transformation $Y = m_1(X)$, and another monotonic transformation $\phi = m_2(\theta)$. It is easy to verify that the diagnostic function for the transformed situation is $D(z, m_2^{-1}(\phi))$; which is to say that $D(z, \theta)$ is invariant under separate monotone transformations of the statistic and the parameter.

Now for the proof of the lemma. The cdf of $\tilde{Z} = q(Z)$ is

$$\tilde{\phi}(\tilde{z}) = \phi(q^{-1}(\tilde{z})) .$$

(2.12)

If $z_\alpha$ is the $100 \alpha^{th}$ normal percentile,
\[ \Phi(z_\alpha) = \alpha , \] (2.13)

the \( \tilde{z}_\alpha = q(z_\alpha) \) is the corresponding percentile of \( \tilde{Z} \). In particular, \( \tilde{Z} \) has median \( q(z, 0.5) = q(0) = 0 \), by (2.6). The density function of \( \tilde{Z} \), \( \tilde{\phi}(z) = \phi'(z) \), satisfies

\[ \tilde{\phi}(z) = \phi(z)/q'(z) , \] (2.14)

so

\[ \tilde{\phi}(0) = \phi(0) = \frac{1}{\sqrt{2\pi}} \] (2.15)

by (2.6).

From (2.5) we get \( F_\theta(x) = \tilde{\phi}\left(\frac{g(x) - \nu_\theta}{\sigma_\theta}\right) \), implying

\[ \tilde{F}_\theta(x) = -\tilde{\phi}\left(\frac{g(x) - \nu_\theta}{\sigma_\theta}\right) \left[ \frac{\nu_\theta}{\sigma_\theta} + \frac{g(x) - \nu_\theta}{\sigma_\theta} \frac{\sigma_\theta}{\sigma_\theta} \right] . \] (2.16)

But, again using (2.5) and the fact that percentiles map in the obvious way,

\[ \tilde{z}_\alpha = \frac{g(x, 0, \theta) - \nu_\theta}{\sigma_\theta} , \] (2.17)

so

\[ \tilde{F}_\theta(x_{\alpha, \theta}) = -\tilde{\phi}(\tilde{z}_\alpha) \left[ \frac{\nu_\theta}{\sigma_\theta} + \frac{\tilde{z}_\alpha}{\sigma_\theta} \frac{\sigma_\theta}{\sigma_\theta} \right] = -\frac{\phi(z_\alpha)}{q'(z_\alpha)} \left[ \frac{\nu_\theta}{\sigma_\theta} + \frac{q(z_\alpha)}{\sigma_\theta} \right] \]

by (2.14). In particular \( \tilde{F}_\theta(x, 0.5, \theta) = -\phi(0) \left[ \frac{\nu_\theta}{\sigma_\theta} \right] \). Substituting these expressions into (2.4) gives (2.7). \( \Box \)

The definition of \( D(z, \theta) \) is motivated in terms of the local transformation to normality, say

\[ t_\theta(x) \equiv \Phi^{-1} F_\theta(x) . \] (2.18)
Under parameter value $\theta$, $t_\theta(X)$ has a $\mathcal{N}(0,1)$ distribution. Definition (2.4) can be rewritten as

$$D(z, \theta) = \frac{t_\theta(x_\theta(z), \theta)}{t_\theta(x, .5, \theta)}.$$  \hspace{1cm} (2.19)

Without going into details, $D(z, \theta)$ measures how quickly the local transformation to normality is changing as $\theta$ varies. (In a Ntf family (2.2), $t_\theta(x) = g(x) - v_\theta$; any local transformation to normality globally normalizes $\mathfrak{J}$ in this case, see Section 5.)

**General Interpretation of $D(z, \theta)$.** The function $D(z, \theta)$ completely determines how the percentiles of the different distributions in $\mathfrak{J}$ relate to one another. In other words, $D(z, \theta)$ determines $\mathfrak{J}$, modulo an arbitrary monotone transformation on the $x$ scale. *These statements hold true for all families $\mathfrak{J}$ having $D(z, \theta)$ defined, not just for families of the G'tf form (2.5).*

The proof of these statements is based on (2.19). First of all, we can assume that

$$\dot{t}_\theta(x, .5, \theta) = -1.$$ \hspace{1cm} (2.20)

If it doesn't, then a change of parameters makes it so: define the new parameter $\xi$ by

$$\frac{d\xi}{d\theta} = \frac{f_\theta(\mu_\theta)}{\phi(0)} \cdot \mu_\theta,$$ \hspace{1cm} (2.21)

where $\mu_\theta = x, .5, \theta$, and $f_\theta(x) = \frac{\partial}{\partial x} F_\theta(x)$. For convenience we assume $\mu_\theta > 0$. It is easy to verify that $\frac{\partial}{\partial \xi} t_\theta(x) \bigg|_{x=\mu_\theta} = -1$. Then (2.20) holds with $\mathfrak{J}$ parameterized by $\xi$ instead of $\theta$. Reparameterizations have no effect on the $D$ function, as commented earlier, nor on the
interpretations of $D$ based on Lemma 1. The $\xi$ parameterization (2.21) is unique (up to an additive constant) for any family $\mathcal{F}$. Here we assume $\theta = \xi$, for ease of notation.

Choose any set of $x$ values, say $x_1, x_2, \ldots, x_k$, and define $z_{\theta} = (t_1(x_1), t_2(x_2), \ldots, t_k(x_k))$. From (2.19) and (2.20) it follows that

$$z_{\theta} = -(D(z_{\theta_1}), D(z_{\theta_2}), \ldots, D(z_{\theta_k})),$$ (2.22)

$z_{\theta_k} = t_k(x_k)$. As $\theta$ moves through $\Theta$, the vector $z_{\theta}$ traces out a curve in $\mathbb{R}^k$ completely describing how the values of $F_\theta(x_k), k=1,\ldots,K$ relate to one another. This curve is determined by the differential equation (2.22), which depends only on the function $D(z,\theta)$.

Suppose that $D(z,\theta) = 1+z \varepsilon_\theta$ as at (2.9). It now follows that $\mathcal{F}$ must be an $N$'tf, without first assuming as we previously did that $\mathcal{F}$ is a $G$'tf. If $D(z,\theta)$ is "nearly" of the form $1+z \varepsilon_\theta$ then $\mathcal{F}$ must "nearly" be an $N$'tf, in a sense made precise by numerically solving (2.22) in any particular case. The same statements hold for the other family types discussed in the next section.

3. Types of Transformation Family. We want to understand how well, or how poorly, a given family $\mathcal{F}$ agrees with the normal transformation form (2.2). To this end it is useful to define more general types of transformation family representing various departures from (2.2). Two such generalizations have already been introduced, the general scaled transformation family (2.5), and the normal scaled transformation family referred to at (2.9).
<table>
<thead>
<tr>
<th>Description</th>
<th>Name and Abbreviation</th>
<th>Relationship</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. $X = g^{-1}(\nu_0 + Z)$</td>
<td>Normal transformation family (N'tf)</td>
<td>N'tf</td>
</tr>
<tr>
<td>2. $X = g^{-1}(\nu_0 + q(Z))$, $q(-z) = -q(z)$</td>
<td>Symmetric transformation family (Stf)</td>
<td>Stf</td>
</tr>
<tr>
<td>3. $X = g^{-1}(\nu_0 + q(Z))$</td>
<td>General transformation family (G'tf)</td>
<td>G'tf</td>
</tr>
<tr>
<td>4. $X = g^{-1}(\nu_0 + \sigma_0 q(Z))$</td>
<td>Normal scaled transformation family (N'tf)</td>
<td>N'tf</td>
</tr>
<tr>
<td>5. $X = g^{-1}(\nu_0 + \sigma_0 q(Z))$, $q(-z) = -q(z)$</td>
<td>Symmetric scaled transformation family (S'tf)</td>
<td>S'tf</td>
</tr>
<tr>
<td>6. $X = g^{-1}(\nu_0 + \sigma_0 q(Z))$</td>
<td>General scaled transformation family (G'tf)</td>
<td>G'tf</td>
</tr>
</tbody>
</table>

Table 1. Six transformation types, described in terms of a standard normal variate $Z$ and the four functions $g^{-1}$, $\nu_0$, $\sigma_0$, and $q$. Arrows in the right diagram indicate increasing generality. Constraints: $q(0) = 0$, $q'(0) = 1$, and $\nu_{\theta_0} = 0$, $\sigma_{\theta_0} = 1$ for some selected $\theta_0$.

Table 1 describes the six types of transformation family used in this paper. The most general case, G'tf, represents a given family $\mathcal{F} = \{X \sim F_{\theta}, \theta \in \Theta\}$ in terms of a standard normal variate $Z$ as follows: $Z$ is transformed to $\tilde{Z} = q(Z)$ by a strictly increasing mapping $q(z)$; $Y = \nu_0 + \sigma_0 \tilde{Z}$ is a scaled and translated version of $\tilde{Z}$; finally $X = g^{-1}(Y)$, where $g^{-1}(y)$ is strictly monotonic. (It is slightly more convenient here to work with $g^{-1}$, rather than with $g$ as in Section 2.) In addition to restrictions (2.6) on $q(z)$, we set

$$\nu_{\theta_0} = 0, \quad \sigma_{\theta_0} = 1 \quad (3.1)$$
for an arbitrary value $\theta_0 \in \Theta$. Sections 4, 5, and 6 show that then the representation $X = g^{-1}(\nu_\theta + \sigma_\theta q(Z))$ is unique, with the exception discussed in Section 4.

Family types 2-5 in Table 1 represent intermediates between the simple N'tf case (2.2) and the G'tf case (2.5). In type 5, for example, S'tf, $\tilde{Z} = q(Z)$ is restricted to be symmetrically distributed about 0. Section 5 shows that some calculations are easier in the S'tf case than in a G'tf.

Figure 3 shows the diagnostic function $D(z, \theta)$ for the case of the normal correlation coefficient. The parameter $\theta = \rho$, the statistic $X = \hat{\rho}$, the sample correlation coefficient. Cramer (1946), Section 29.7, describes the distribution of $X$ as a function of $\theta \in (-1, 1)$.

Figure 3. The diagnostic function $D(z, \theta)$ for the normal correlation coefficient, $\theta = \rho$, $X = \hat{\rho}$. The upper curves are for sample size $n = 15$, the lower for $n = 5$. 

14
The upper set of curves applies to \( n \), the number of pairs of bivariate normal points going into the computation of \( \hat{\beta} \), equal to 15. We shall see that this is quite nearly a symmetric transformation family, \( Stf \), with \( \tilde{Z} = q(Z) \) a Student's \( t \) variate with 38 degrees of freedom, Section 4. This is as close as we come in this paper to a genuine example of a normal transformation family.

The lower set of curves apply to \( n = 5 \). This turns out to be close to an \( S'tf \) situation, with \( \tilde{Z} = q(Z) \) a Student's \( t \) variate with 10 degrees of freedom. The scale function \( \sigma_0 \) has its maximum at \( \theta = \rho = 0 \), and decreases as \( \theta \) goes to \( +1 \): at \( \theta = .5 \), \( \varepsilon_0 = \dot{\sigma}_0 / \nu_0 \) equals \(-.041\); at \( \theta = .9 \) it equals \(-.060\).

If we assume that the normal correlation family \( \mathcal{F} \) is a \( G'tf \), then it must be a \( S'tf \). This follows from the symmetry of \( \mathcal{F} \) about \( 0 \), i.e. the fact that the mapping \( (\theta, X) \rightarrow (-\theta, -X) \) takes \( \mathcal{F} \) into itself, and the uniqueness of the \( G'tf \) representation. The proof will not be given here. The same considerations apply to the binomial family, Section 7.

The three types \( Ntf, Stf, Gtf \) are the most useful for statistical applications. In these cases there exists an obvious 1-2 central confidence interval for \( \theta \) based on observing \( Y = g(X) \),

\[
\nu_0 \in [y-z_{1-\alpha}, y-z_{\alpha}],
\]

where \( \tilde{z}_\alpha \) is the 100 \( \alpha \)th percentile of \( \tilde{Z} = q(Z) \). This transforms back into a confidence interval for \( \theta \), assuming we know the mapping from \( \nu_0 \) back to \( \theta \). The usual approximate intervals for the normal correlation coefficient are based on this device.

In a \( Gtf \), and therefore also in an \( Ntf \) or \( Stf \), \( \sigma_0 \equiv 1 \), so \( \varepsilon_0 = \dot{\sigma}_0 / \nu_0 = 0 \) and the lemma gives
\[ D(z, \theta) = \frac{1}{q'(z)} \quad \text{(3.3)} \]

not depending upon \( \theta \). It will turn out that \( D(z, \theta) \) not depending upon \( \theta \) is a sufficient as well as necessary condition for the simple types Ntf, Stf, Gtf, Section 4.

4. Finding \( q(Z) \). The representation \( X = g^{-1}(v_\theta + \sigma_\theta q(Z)) \) involves four functions, \( g, v_\theta, \sigma_\theta, \) and \( q \). We will give simple methods for calculating these functions directly from the cdf's \( F_\theta(x) \). This section concerns the calculation of \( q(z) \) which can be done in terms of the diagnostic function \( (2.4) \). For convenient discussion, henceforth "G'Gf" includes any of the six family types in Table 1, and "Gtf" includes the three types Ntf, Stf, Gtf. In other words, any type name refers to its description in Table 1, and also to all more restrictive types in that table.

**Theorem 1.** Suppose the family \( \mathcal{F} \) is a G'tf, \( X = g^{-1}(v_\theta + \sigma_\theta q(Z)) \), and that there exists two values \( \theta_1 \) and \( \theta_2 \) such that \( D(z, \theta_1) \) is not identically equal to \( D(z, \theta_2) \). Then

\[ \frac{d}{dz} \log(q(z)/z) = \frac{D'(0, \theta_2) - D'(0, \theta_1)}{D(z, \theta_2) - D(z, \theta_1)} - \frac{1}{z}, \quad \text{(4.1)} \]

where \( D'(0, \theta) = \frac{d}{dz} D(z, \theta) \Big|_{z=0} \). If \( D(z, \theta) \) does not depend on \( \theta \), then \( \mathcal{F} \) is a Gtf, \( X = g^{-1}(v_\theta + q(Z)) \), and

\[ q'(z) = \frac{1}{D(z)}. \quad \text{(4.2)} \]

**Proof.** From Lemma 1 and \((2.7)\) we get

\[ D(z, \theta_2) - D(z, \theta_1) = (\epsilon_{\theta_2} - \epsilon_{\theta_1}) \frac{q(z)}{q'(z)} \quad \text{(4.3)} \]
for any $\theta_1, \theta_2 \in \Theta$. Also, using (2.6),

$$D'(0, \theta) = \frac{3}{\partial z} \left[ \frac{1 + q(z) \varepsilon_\theta}{q'(z)} \right]_{z=0} = -q''(0) + \varepsilon_\theta.$$ (4.4)

Notice that $D(z, \theta_2)$ not being identically equal to $D(z, \theta_1)$ is equivalent to $\varepsilon_{\theta_2} \neq \varepsilon_{\theta_1}$. Assuming this is the case, (4.3) and (4.4) give

$$\frac{D'(0, \theta_2) - D'(0, \theta_1)}{D(z, \theta_2) - D(z, \theta_1)} - \frac{1}{z} = \frac{q'(z)}{q(z)} - \frac{1}{z} = \frac{d}{dz} \log \frac{q(z)}{z},$$ (4.5)

which is (4.1).

Next suppose that $D(z, \theta)$ does not depend on $\theta$. From (4.3) we see that this is possible only if $\varepsilon_\theta = \varepsilon_\sigma_\theta = (d \sigma / d \nu)_\theta$ is a constant, i.e. if

$$\sigma_\theta = 1 + cv_\theta$$ (4.6)

for some constant $c$. (The intercept equals 1 because of (3.1).) In this case we can rewrite the G'tf representation $g(X) = v_\theta + \sigma_\theta \tilde{z}$ as

$$1 + cg(X) = (1 + cv_\theta) + (1 + cv_\theta)c\tilde{z} = (1 + cv_\theta)(1 + c\tilde{z}).$$ (4.7)

Letting $g^\circ(x) = \frac{1}{c} \log(1 + cg(x))$, $v_\theta^\circ = \frac{1}{c} \log(1 + cv_\theta)$, and $\tilde{z}^\circ = \frac{1}{c} \log(1 + c\tilde{z})$ gives

$$g^\circ(X) = v^\circ_\theta + \tilde{z}^\circ,$$ (4.8)

a G'tf representation. (Notice that $q^\circ(Z) = \frac{1}{c} \log(1 + c\tilde{z}) = \frac{1}{c} \log(1 + cq(Z))$ satisfies (2.6). The difficulty with $1 + c\tilde{z}$ possibly being negative is discussed below.) But in a G'tf we have $\sigma_\theta \equiv 1$ so $\varepsilon_\theta \equiv 0$, and $D(z, \theta) = 1/q'(z)$ by (2.7), giving (4.2). □
In the G'tf case, the theorem allows us to calculate \( \frac{d}{dz} \log \frac{q(z)}{z} \) from \( D(z, \theta_1), D(z, \theta_2) \). This gives \( q(z)/z \) up to a multiplicative constant, whose value is then determined by the condition \( \lim_{z \to \infty} q(z)/z = 1 \), derived from (2.6). The main point is that \( q(z) \) is determined directly from \( D(z, \theta) \), and therefore must be a unique function of the family of cdf's \( F_\theta(x) \). Having obtained \( q(z) \), the functions \( g, \nu_\theta, \) and \( \sigma_\theta \) are also uniquely determined by the cdf's of \( X \), see Sections 5 and 6.

Uniqueness breaks down in the Gtf case. The Gtf representation \( g(X) = \nu_\theta + \tilde{z} = \nu_\theta + q(z) \) can be rewritten as

\[
g^o(X) = \nu_\theta^o + (1 + c\nu_\theta^o)q^o(z),
\]

where \( c \) is any constant and \( g^o(x) = \{\exp[cg(x)] - 1\}/c, \nu_\theta^o = \{\exp[c\nu_\theta]\} - 1\}/c, \) and \( q^o(z) = \{\exp[cq(z)] - 1\}/c. \) Representation (4.9) is a G'tf with \( \epsilon_\theta = \sigma_\theta^o/\nu_\theta^o = c. \) In other words, corresponding to any Gtf is a one-parameter family of G'tf representations, the free parameter being the constant value of \( \epsilon_\theta \). The Gtf representation, having \( \epsilon_\theta \equiv 0, \) is obtained from (4.2). There is only one such representation for a given Gtf family \( J \), and so the uniqueness of the representation theory continues to hold if we agree to always represent Gtf's as such.

We can work backwards and ask how a certain form of \( q(z) \) affects the \( D(z, \theta) \) function. Suppose that we know we are in an S'tf situation so that \( q(-z) = -q(z) \). Writing

\[
q(z) = z + \frac{Bz^3}{6} + \ldots,
\]

(4.10) gives

\[
D(z, \theta) = 1 + \epsilon_\theta z - \frac{B}{3} z^2 \ldots.
\]

(4.11)
Stopping the series after the quadratic term gives a reasonable approximation to \( D(z, \theta) \) for the normal correlation coefficient, Figure 3.

For \( n = 5 \), \( B \approx .146 \), for \( n = 15 \), \( B \approx .039 \).

The Cornish-Fisher expansion for a Student-t variate with \( d \) degrees of freedom, rescaled to have the same density as a \( \mathcal{N}(0,1) \) at \( z=0 \), begins \( q(z) = z + z^3/(4d+2) \), Johnson and Kotz (1970), p. 102. Comparing this with (4.10) gives the approximation

\[
d \approx \frac{1}{4} \left( \frac{6}{B} - 2 \right).
\]  

(4.12)

This gives \( d \approx 10 \) for \( n = 5 \) and \( d \approx 38 \) for \( n = 15 \) in the case of the normal correlation coefficient.

We interpreted Figure 2, for the continuous Poisson case, as if \( D(z, \theta) \) were linear in \( z \). In fact, \( D(z, \theta) \) displays a small amount of curvature, which is particularly evident for \( \theta = 1 \). Just how small this curvature is can be seen by comparison with (4.10), (4.11). The maximum possible value of \( B \) in Figure 2 is about .025, giving \( d = 60 \). For almost any purpose, a \( t_{60} \) variate is an excellent approximation to a \( \mathcal{N}(0,1) \) variate, so the interpretation of the continuous Poisson family as an \( N'tf \) seems quite reasonable.

In an \( S'tf \) we have the simple relationship

\[
\varepsilon_\theta = D'(0, \theta) \quad \text{(in an } S'tf),
\]

(4.13)

so that \( \varepsilon_\theta \) can be read directly from the graph of \( D(z, \theta) \). This follows from (2.7), which gives

\[
D'(0, \theta) = \frac{q'(0)^2 \varepsilon_\theta - (1+\varepsilon_\theta q(0))q''(0)}{q'(0)^2} = \varepsilon_\theta - q''(0).
\]

(4.14)
In an S'f, $q''(0) = 0$ by symmetry, giving (4.13).

In going from (4.7) to (4.8) we ignored the possibility $1 + \tilde{c} \tilde{Z} < 0$. The following special case illustrates what happens in this situation. Let $\mathcal{X}$ be the family

$$X \sim h(\theta, (1+\varepsilon \theta)^2) \quad (\theta > -1/\varepsilon).$$

(4.15)

Here $\varepsilon$ is a positive constant. This family is an N'tf, $q(z) = z$, with $g$ the identity mapping, $v_\theta = \theta$, $\sigma_\theta = 1+\varepsilon \theta$. By (2.9), $D(z, \theta) = 1+\varepsilon z$.

We can also write (4.15) as

$$(1+\varepsilon X) = (1+\varepsilon \theta)(1+\varepsilon Z),$$

(4.16)

$Z \sim h(0,1)$. Notice that the sign of $1+\varepsilon X$ is an ancillary statistic, $\text{Prob}(1+\varepsilon X > 0) = \text{Prob}(1+\varepsilon Z > 0) = \Phi(1/\varepsilon)$ independent of $\theta$. We can separately transform the two sign cases,

$$\frac{\log(1+\varepsilon X)}{\varepsilon} = \frac{\log(1+\varepsilon \theta)}{\varepsilon} + \frac{\log(1+\varepsilon Z)}{\varepsilon} \quad \text{(for } 1+\varepsilon X > 0)$$

and

$$\frac{\log-(1+\varepsilon X)}{\varepsilon} = \frac{\log(1+\varepsilon \theta)}{\varepsilon} + \frac{\log-(1+\varepsilon Z)}{\varepsilon} \quad \text{(for } 1+\varepsilon X < 0).$$

(4.17)

In other words, (4.15) is a Gtf "on two real lines", one real line corresponding to each sign of $1+\varepsilon X$. The parameter $v_\theta^\circ = \frac{1}{\varepsilon} \log (1+\varepsilon \theta)$ translates the distribution of $X^\circ = \frac{1}{\varepsilon} \log|1+\varepsilon X|$ in the usual way, except that there is always total probability $\Phi(-1/\varepsilon)$ on the line corresponding to $1+\varepsilon X < 0$, and total probability $\Phi(1/\varepsilon)$ on the line corresponding to $1+\varepsilon X > 0$. Probability cannot move from one line to the other, no matter how $v_\theta^\circ$ varies. Formula (4.2),

20
\[ q'(z) = \frac{1}{1 + \varepsilon z} \quad (4.18) \]

in this case, gives both transformations of \( Z \) in (4.17).

To summarize, if \( D(z,\theta) \) does not depend on \( \theta \) then \( \mathfrak{F} \) is a Gtf, though possibly defined on two lines. Formula (4.12) gives the complete solution of \( q(z) \).

Formula (4.1) has been written in a form convenient for numerical computation. Other expressions are possible, for example

\[ \frac{d}{dz} \log |q(z)| = \frac{\partial}{\partial \theta} \frac{D'(0,\theta)}{D(z,\theta)} \quad (4.19) \]

In a G'tf (4.1), or (4.19), does not depend on the choice of \( \theta \) values. If this is markedly untrue then \( \mathfrak{F} \) does not have a good G'tf approximation.

It is easy to see when a G'tf family \( \mathfrak{F} \) is actually S'tf. The function \( D_+(z,\theta) \equiv D(z,\theta) + D(-z,\theta) \) equals

\[ D_+(z,\theta) = \frac{2}{q'(z)} + \varepsilon_\theta \frac{q(z) + q(-z)}{q'(z)} \quad (4.20) \]

in a G'tf. Assuming that \( \varepsilon_\theta \) is not constant, i.e. that \( \mathfrak{F} \) is not a Gtf family, then \( D_+(z,\theta) \) not depending on \( \theta \) is necessary and sufficient for \( \mathfrak{F} \) to be S'tf, since both conditions are equivalent to \( q(z) + q(-z) = 0 \). If \( \mathfrak{F} \) is a Gtf then (4.2) gives \( D(z) = D(-z) \) as necessary and sufficient for \( \mathfrak{F} \) to be Stf.

21
5. Finding \( g(X) \). We wish to compute the function \( g \) in the G'lf representation \( g(X) = \nu_1 + \sigma_\theta q(z) \). Two formulas will be given, one for the general G'lf case and a simpler one applying to the Gtf case. Let \( x_1 < x_2 \) be any two values of \( X \), and define \( \theta_{12} \) as that value of \( \theta \) making \( F_{\theta}(x_1) = 1 - F_{\theta}(x_2) \), say

\[
\alpha = F_{\theta_{12}}(x_1) = 1 - F_{\theta_{12}}(x_2) .
\]  

(5.1)

Also define \( \mu^{-1}(x) \) as that value of \( \theta \) such that \( x \) is the median of \( X \),

\[
F_{\mu^{-1}(x)}(x) = 0.5 ,
\]  

(5.2)

and let

\[
f_{\theta}(x) = F'_{\theta}(x) ,
\]  

(5.3)

be the density function of \( X \). (The prime will always indicate differentiation with respect to the argument in parentheses.)

**Theorem 2.** In a G'lf family,

\[
\frac{g'(x_2)}{g'(x_1)} = \frac{f_{\theta}(x_2)}{f_{\theta}(x_1)} \frac{q'(z_{1-\alpha})}{q'(z_{\alpha})} .
\]  

(5.4)

The simpler formula

\[
\tilde{g}'(x) = \frac{f_{\mu^{-1}(x)}}{\phi(0)}
\]  

(5.5)

gives \( g'(x) \) in Gtf family.
Proof. Since \( F_\theta(x) = \phi\left(\frac{g(t) - v_\theta}{\sigma_\theta}\right) \) where \( \phi(z) = \phi(q^{-1}(z)) \) as at (2.12), differentiation yields \( f_\theta(x) = \phi\left(\frac{g(x) - v_\theta}{\sigma_\theta}\right) g'(x) \). Substituting \( x = x_\alpha, \theta \) gives
\[
f_\theta(x_\alpha, \theta) = \phi(z_\alpha) \frac{g'(x_\alpha, \theta)}{\sigma_\theta} = \frac{\phi(z_\alpha)g'(x_\alpha, \theta)}{q'(z_\alpha)\sigma_\theta},
\]
the last equality following from (2.14). But for \( \theta = \theta_{12} \), we have \( x_1 = x_\alpha, \theta, x_2 = x_{1-\alpha}, \theta \), and \( \phi(z_\alpha) = \phi(z_{1-\alpha}) \), so (5.6) follows from (5.4) by division.

The Gtf formula (5.5) follows from the last equality in (5.6). We take \( \theta = \mu^{-1}(x), \alpha = .5 \), so \( x = x_\alpha, \theta \) and \( z_\alpha = 0 \). Since \( \sigma_\theta = 1 \) in a Gtf, and \( q'(0) = 1 \) by (2.6), (5.6) gives \( g'(x) = \frac{f_{\mu^{-1}(x)}}{g'(x)} \) as claimed.

Formula (5.4) simplifies in the S'tf situation to
\[
\frac{g'(x_2)}{g'(x_1)} = \frac{f_{\theta_{12}}(x_2)}{f_{\theta_{12}}(x_1)},
\]
since \( q'(z_{1-\alpha}) = q'(z_\alpha) \) by symmetry. Like (5.5), formula (5.7) has the advantage of not requiring knowledge of \( q(z) \).

Formula (5.4), and (5.7) in the S'tf case, are convenient for numerical computation, as demonstrated in Section 7 where we consider the Poisson and binomial cases. From a starting value \( x_1 \), and with \( \alpha \) fixed, we calculate \( x_2, x_3, x_4, \ldots \) and \( \theta_{12}, \theta_{23}, \theta_{34}, \ldots \) satisfying \( \alpha = F_{\theta_{i-1,i-1}}(x_{i-1}) = 1 - F_{\theta_{i-1,i}}(x_i) \). Successive use of (5.4) gives \( g'(x_1), g'(x_2), g'(x_3), \ldots \) up to a multiplicative constant. This leaves two degrees of freedom in the determination of \( g(x) \), a multiplicative and an additive constant, which are determined by (3.1), as shown in Section 6, expression (6.4).
Letting \( \alpha = 0.5 \) in (5.4) gives a "single x" version,

\[
\frac{d}{dx} \log g'(x) = \frac{d}{dx} \log f_\theta(x) \bigg|_{\theta = \mu^{-1}(x)} + q''(0) \frac{\mu^{-1}(x)}{\phi(0)}. \tag{5.8}
\]

The last term vanishes in an S'tf, since \( q''(0) = 0 \).

Formula (5.5) has a simple intuitive interpretation in terms of the local transformation to normality \( t_\theta(x) = \phi^{-1} F_\theta(x) \), (2.18). In an Ntf,
\[X = g^{-1}(\nu_\theta + Z),\]
we have \( t_\theta^\prime(\nu_\theta) = g'(x) - \nu_\theta \), so

\[
t_\theta^\prime(\nu_\theta) = g'(x). \tag{5.9}
\]

This means it doesn't matter which value of \( \theta_0 \) we choose: \( t_\theta^\prime(x) \) always agrees with \( g'(x) \) in an Ntf. In other words, any local transformation to normality globally normalizes an Ntf.

If \( \mathcal{F} \) is not an Ntf then (5.9) doesn't hold. However, we can try to choose among the different \( t_\theta(x) \) transformations by selecting that \( \theta \) most appropriate to each \( x \). An obvious choice is \( \theta = \mu^{-1}(x) \), with \( t_\theta(x) \) having \( x \) derivative

\[
t_\theta^\prime(\mu^{-1}(x)) = \frac{f(\mu^{-1}(x))}{\phi(0)}. \tag{5.10}
\]

which is formula (5.5). In words, \( \tilde{g}(x) \) is the transformation everywhere having the same \( x \) derivative as \( t_\theta(x) \), evaluated at that \( \theta \) for which \( x \) is the median of \( X \).

In an important sense \( \tilde{g}(x) \) deserves to be called a variance stabilizing transformation. In a Gtf, where perfect stabilization is possible, \( \tilde{g}(x) \) achieves this exactly:

\[Y = \tilde{g}(X) = \nu_\theta + Z,\]
a translation family, with constant variance.
The following corollary shows that $\tilde{g}(x)$ tries to stabilize variances in the more general context of a G'tf.

**Corollary 1.** If $\mathcal{G}$ is a G'tf, $X \sim g^{-1}(\mathcal{G}_{\mathcal{G}}+\sigma_{\mathcal{G}}^2 Z)$, then

$$\tilde{g}'(x) = \frac{g'(x)}{\sigma_{\mathcal{G}}^2} \left( \frac{\mathcal{G}_{\mathcal{G}}}{\mu^{-1}(x)} \right).$$  \hspace{1cm} (5.11)

**Proof.** Taking $\alpha = .5$ in (5.6) gives

$$f = \frac{g'(x)}{\sigma_{\mathcal{G}}^2} \frac{\mathcal{G}_{\mathcal{G}}}{\mu^{-1}(x)} \phi(0),$$  \hspace{1cm} (5.12)

and the corollary follows immediately from definition (5.5). \(\square\)

Here is the interpretation of (5.11): first make the transformation $Y = g(X)$, which produces a location scale family $Y = \mathcal{G}_{\mathcal{G}}+\sigma_{\mathcal{G}}^2 Z$. Now apply (5.5) to this family. Since $Y$ has density $1/\sigma_{\mathcal{G}}^2 \phi((y-\mathcal{G}_{\mathcal{G}})/\sigma_{\mathcal{G}}^2)$, (5.5) gives the transformation of $Y$, call it $h(y)$, with derivative at $y = g(x)$ equal to

$$h'(y) = \frac{1/\sigma_{\mathcal{G}}^2}{\mathcal{G}_{\mathcal{G}}(y)} = \frac{1}{\sigma_{\mathcal{G}}^2} \frac{1}{\mu^{-1}(x)}.$$  \hspace{1cm} (5.13)

(Here we have used (2.15), and the fact that if $\text{Prob}_{\mathcal{G}}(X<x) = .5$, that is if $\theta = \mu^{-1}(x)$, then $\text{Prob}_{\mathcal{G}}(Y<y) = .5$, that is $\theta = \nu^{-1}(y)$.) According to (5.11), the transformation $\tilde{g}(x)$ is the composition $hg(x)$.

In the case of an N'tf, $X = g^{-1}(\mathcal{G}_{\mathcal{G}}+\sigma_{\mathcal{G}}^2 Z)$, the transformed variable $Y = g(X)$ is perfectly normal, $Y \sim N(\theta, \sigma_{\mathcal{G}}^2)$, but with nonconstant variance. Then (5.5) makes the further transformation $W = h(Y)$, where $h'(y) = 1/\sigma_{\mathcal{G}}^2$, which spoils the normality but tends to produce more constant variance. Section 7 discusses the tradeoff between normality and constant
variance for the Poisson family. Section 8 concerns the relative merits of stabilization versus normalization.

6. Finding \( \nu_\theta \) and \( \sigma_\theta \). Having found \( g \) and \( q \) in the G'tf representation \( X = g^{-1}(\nu_\theta + \sigma_\theta q(Z)) \), it is easy to compute the location and scale functions \( \nu_\theta, \sigma_\theta \). Define \( \mu_\theta \) to be the median of \( X \) for parameter value \( \theta \),

\[
\mu_\theta: \quad F_\theta(\mu_\theta) = .5 .
\]  \hspace{1cm} (6.1)

Notice that the function \( \mu_\theta \) can be computed directly from the cdf's \( F_\theta \) comprising \( \mathcal{J} \), without any knowledge of the G'tf representation. (This doesn't mean it is easy to find a formula for \( \mu_\theta \). In our examples the computations are done numerically.)

Because \( g(X) \) is a monotonic mapping, and because \( \nu_\theta \) is the median of \( g(X) = \nu_\theta + \sigma_\theta q(Z) \), by (2.6), we have

\[
\nu_\theta = g(\mu_\theta) .
\]  \hspace{1cm} (6.2)

This is an obvious formula, of course, but it is often overlooked in the literature, where there is a tendency to automatically take \( g(\theta) \) as the center for the distribution of \( g(X) \).

The scale function \( \sigma_\theta \) is computed from (5.12),

\[
\sigma_\theta = \frac{\phi(0) \ g'(\mu_\theta)}{f_\theta(\mu_\theta)} .
\]  \hspace{1cm} (6.3)

Notice that (3.1) can now be rewritten as

\[
g(\mu_{\theta_0}) = 0, \quad g'(\mu_{\theta_0}) = \frac{f_{\theta_0}(\mu_{\theta_0})}{\phi(0)}. 
\]  \hspace{1cm} (6.4)
These two constraints complete the determination of the function $g$ from formula (5.4). Comparing (6.4) with (5.5) shows that $g'(\mu_{\theta_0}) = \hat{g}'(\mu_{\theta_0})$.

For two values $\theta_1, \theta_2$, let $x_1 = \mu_{\theta_1}$ and $x_2 = \mu_{\theta_2}$. Then (5.4) and (6.3) together give

$$\frac{\sigma_{\theta_2}}{\sigma_{\theta_1}} = \frac{f_{\theta_{12}}(x_2)}{f_{\theta_{12}}(x_1)} \frac{f_{\theta_1}(x_1)}{f_{\theta_1}(x_2)} \frac{q'(z_{1-y})}{q'(z_{\alpha})}. \quad (6.5)$$

In an S'tf family, where $q'(-z) = q'(z)$, this reduces to

$$\frac{\sigma_{\theta_2}}{\sigma_{\theta_1}} = \frac{f_{\theta_{12}}(x_2)}{f_{\theta_{12}}(x_1)} \frac{f_{\theta_1}(x_1)}{f_{\theta_2}(x_2)}, \quad (6.6)$$

which is convenient for computation, especially in conjunction with (5.7).

**Note.** Formulas (6.2), (6.3) for $\nu_\theta, \sigma_\theta$ involve $g$ but not $q$. This can be reversed. It is fairly obvious that once $q(z)$ is known, we can determine $\sigma_{\theta_2}/\sigma_{\theta_1}$ and $(\nu_{\theta_2} - \nu_{\theta_1})/\sigma_{\theta_1}$ for any two parameter values $\theta_1, \theta_2$, simply by comparing $F_{\theta_1}(x)$ with $F_{\theta_2}(x)$ at different values of $x$, as in (1.7). Combined with (3.1), this gives $\sigma_\theta$ and $\nu_\theta$. An example appears in Section 7.

7. **The Continuous Poisson Family.** Figure 2 shows that the continuous Poisson family $\mathcal{G}$ is nearly an N'tf, $X = g^{-1}(\nu_\theta + \sigma_\theta Z)$. Now we use the formulas of Sections 5 and 6 to calculate the functions $g$, $\nu_\theta$, and $\sigma_\theta$, and also the variance stabilizing transformation $\hat{g}$, for $\mathcal{G}$. The results are shown in Table 2.
<table>
<thead>
<tr>
<th>x</th>
<th>.125</th>
<th>.250</th>
<th>.5</th>
<th>1</th>
<th>2</th>
<th>4</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>θ = μ⁻¹(x)</td>
<td>.334</td>
<td>.450</td>
<td>.691</td>
<td>1.180</td>
<td>2.164</td>
<td>4.161</td>
<td>8.159</td>
</tr>
<tr>
<td>σ₀</td>
<td>.85</td>
<td>.88</td>
<td>.93</td>
<td>1</td>
<td>1.09</td>
<td>1.19</td>
<td>1.29</td>
</tr>
<tr>
<td>ν₀</td>
<td>-1.00</td>
<td>-.85</td>
<td>-.53</td>
<td>0</td>
<td>.85</td>
<td>2.14</td>
<td>4.15</td>
</tr>
<tr>
<td>g'(x)</td>
<td>1.45</td>
<td>1.29</td>
<td>1.12</td>
<td>.93</td>
<td>.75</td>
<td>.58</td>
<td>.46</td>
</tr>
<tr>
<td>(c₀ x⁻¹/₃)</td>
<td>(1.86)</td>
<td>(1.47)</td>
<td>(1.17)</td>
<td>(.93)</td>
<td>(.74)</td>
<td>(.59)</td>
<td>(.47)</td>
</tr>
<tr>
<td>g'(x)</td>
<td>1.58</td>
<td>1.40</td>
<td>1.18</td>
<td>.93</td>
<td>.69</td>
<td>.49</td>
<td>.35</td>
</tr>
<tr>
<td>(c₁(x+.33)⁻¹/₂)</td>
<td>(1.59)</td>
<td>(1.41)</td>
<td>(1.18)</td>
<td>(.93)</td>
<td>(.70)</td>
<td>(.52)</td>
<td>(.37)</td>
</tr>
<tr>
<td>(c₂(x+.375)⁻¹/₂)</td>
<td>(1.54)</td>
<td>(1.38)</td>
<td>(1.17)</td>
<td>(.93)</td>
<td>(.71)</td>
<td>(.52)</td>
<td>(.38)</td>
</tr>
</tbody>
</table>

Table 2. Continuous Poisson family (2.10). The functions g, ν₀, and σ₀ are calculated for the representation X = g⁻¹(ν₀ + σ₀Z), under the constraints ν₁.180 = 0, σ₁.180 = 1; also calculated is the variance stabilizing transformation g’. Figures in parentheses relate to the traditional normalizing and variance stabilizing transformations. The constants c₀ = .93, c₁ = 1.07, c₂ = 1.09 are included to make the derivatives equal .93 at x = 1.
In Table 2, constraint (3.1) was applied with \( \theta_0 = \mu^{-1}(1) = 1.180 \); (6.4) then gives \( g'(1) = .93 \). The function \( g'(x) \) was obtained from (5.7). (The stepwise algorithm described in the paragraph following (5.7) was applied with \( \alpha = .45 \).) The functions \( v_\theta \) and \( \sigma_\theta \) were obtained from (6.2) and (6.3), respectively. Formula (5.5) gave \( \bar{g}'(x) \), also constrained to have \( \bar{g}'(1) = .93 \). Note: These calculations are valid assuming that \( \bar{g} \) is an S'tf, not necessarily an N'tf.

Anscombe (1953) suggested \( x^{2/3} \) as a normalizing transformation for the Poisson, on the grounds that this transformation makes the skewness approximately zero. The derivative \( x^{-1/3} \), suitably rescaled, is seen to agree well with \( g'(x) \) for \( x \leq 1 \), but, perhaps unsurprisingly, not for \( x < 1 \).

The variance stabilizing transformation \( (x + .375)^{1/2} \), Anscombe (1948), has its derivative agreeing well with \( \bar{g}'(x) \) over the entire range of \( x \). The best agreement with \( \bar{g}'(x) \) among functions of the form \( (x+b)^{-1/2} \) is obtained for \( b = .33 \). As a matter of fact, \( (x + .33)^{1/2} \) stabilizes variances within the genuine Poisson family just as well as does \( (x + .375)^{1/2} \), both transformations being superior to the naive transformation \( x^{1/2} \) in this regard. In this case, formula (5.5) has produced an excellent variance stabilizer.

Looking at Table 2, we see that \( \bar{g}(x) \) is a more extreme transformation than \( g(x) \), its derivative being everywhere more quickly varying. A natural measure of this increased "strength of transformation", cf. Tukey (1957), is

\[
\frac{d}{dx} \frac{\log \bar{g}'(x)}{\log g'(x)}. \tag{7.1}
\]
Quantity (7.1) equals approximately 1.37 over the entire range of \( x \) in Table 2. We can state the situation for the continuous Poisson family as follows: There is a transformation \( g(x) \) which nearly normalizes \( \mathcal{F} \), but in order to stabilize variances we must everywhere increase the strength of this transformation by about 37%.

The Genuine Poisson Family. One might worry that our description of the continuous Poisson family was irrelevant to the genuine Poisson family. It is easy to allay such fears. We now argue that any family \( \mathcal{F} \) of continuous distributions which agrees with the Poisson family at the half-integer points, as at (2.11), must give similar results.

For \( j \) and \( k \) nonnegative integers, define \( \theta = \mu^{-1}(j + \frac{1}{2}) \), 

\[
z = \phi^{-1} F_\theta(k + \frac{1}{2}).
\]

Then

\[
D(z, \theta) = \frac{F_\theta(k + \frac{1}{2})}{F_\theta(j + \frac{1}{2})} \frac{\phi(0)}{\phi(1)}
\]

(7.2)

is determined by the behavior of \( F_\theta(j + \frac{1}{2}) \) and \( F_\theta(k + \frac{1}{2}) \) as functions of \( \theta \). Since \( \mathcal{F} \) agrees with the genuine Poisson family at \( j + \frac{1}{2} \) and \( k + \frac{1}{2} \), the value of \( D(z, \theta) \) must be the same as for the continuous Poisson for these \( (z, \theta) \) combinations. The same argument shows that \( D(z_2, \theta)/D(z_1, \theta) \) agrees with its value in the continuous Poisson family, for any \( \theta \) and for \( z_1 = \phi^{-1} F_\theta(k_1 + \frac{1}{2}), z_2 = \phi^{-1} F_\theta(k_2 + \frac{1}{2}) \), where \( k_1, k_2 \) are nonnegative integers.

We would only be interested in \( \mathcal{F} \) if it had a simpler representation than the continuous Poisson family. In particular, we might hope that \( \mathcal{F} \) could be represented as an Ntf rather than an N'tf. Comparing Figure 1 with Figure 4, this is seen to be impossible. Assuming that \( \mathcal{F} \)
Figure 4. For any family $\mathcal{F}$ agreeing with the genuine Poisson family at half-integer points, $D(z,\theta)$ has the same value as in the continuous Poisson family at the indicated points $(z,\theta)$. Only points with $|z| < 2$, $\theta < 4$, are indicated. The ratio $D(z_2,\theta)/D(z_1,\theta)$ agrees with its value in the continuous Poisson family for $z_1$ and $z_2$ on the indicated curves, for any value of $\theta$.

is nearly an N'tf, it is clear that $D(z,\theta)$ must agree closely with Figure 1. In particular, $\epsilon_\theta = \sigma_\theta/\nu_\theta$ must be nearly as shown in the inset.

Similar arguments can be given that $g$, $\nu_\theta$, and $\sigma_\theta$ must be nearly as shown in Table 2, if $\mathcal{F}$ agrees with the Poisson as at (2.11). For example, let $z_{\theta k} = \Phi^{-1} F_\theta(k + \frac{1}{2})$ for $\theta = 1, 2$ and $k=0, 1, 2, \ldots$. 

31
Making use of a table of Poisson probabilities, the pairs \((z_{1k}, z_{2k})\)
equal \((-0.34, -1.10), (0.63, -0.24), (1.40, 0.46), (2.08, 1.07), (2.68, 1.62)\)
for \(k=0, 1, 2, 3, 4\). If \(\mathcal{F}\) is an N'tf, \(F_\theta(x) = \Phi\left(\frac{g(x)-\nu_\theta}{\sigma_\theta}\right)\), then we
have the linear relationship

\[
z_{2k} = \frac{\sigma_1}{\sigma_2} z_{1k} - \frac{\nu_2 - \nu_1}{\sigma_2}.
\]  
(7.3)

as at (1.7). A plot of the five pairs \((z_{1k}, z_{2k})\) given above shows near
perfect linearity, with \(\sigma_1/\sigma_2 = .90\), \((\nu_2 - \nu_1)/\sigma_2 = .81\). Starting with
\(\sigma_1 = 1, \nu_1 = 0\), say, and going on in this way we can recover \(\sigma_\theta\) and \(\nu_\theta\)
for all positive integer values of \(\theta\). These calculations only use the
genuine Poisson probabilities \(F_\theta(k + \frac{1}{2})\).

The story for the binomial family \(\text{Bi}(n, \theta)\), \(n\) fixed, is similar to
that for the Poisson. Figure 5 shows the diagnostic function \(D(z, \theta)\) for
the continuous Binomial family

\[
F_\theta(x) = \int_0^1 t^{x-0.5} (1-t)^{n-x-0.5} \frac{\Gamma(n+1)}{\Gamma(x+0.5) \Gamma(n-x+0.5)} dt,
\]  
(7.4)

\((-\frac{1}{2} < x < n + \frac{1}{2})\),

\(\theta \in (0, 1), n = 20\). The cdf \(F_\theta(x)\) equals that for the genuine binomial
at the half-integer points \(1/2, 3/2, \ldots, (n-1)/2\). Again we have nearly
an N'tf family, with \(\epsilon_\theta\) large for \(\theta\) near 0 or 1. The equivalent
of Table 2 will not be presented here.
8. **Normalization Versus Stabilized Variance.** Suppose that there exists a monotonic transformation \( g(x) \) such that \( g(X) \) is (nearly) normally distributed for every value of the parameter \( \theta \). That is, suppose that \( X \) is (nearly) an \( \text{N}'\text{t}f \), as is the continuous Poisson family. The point of this section is that under certain circumstances we might still prefer to work with the variance stabilizing transformation \( \tilde{g}(X) \), (5.5). These calculations are far from conclusive. They are intended only as a cautionary note against uncritical use of normality as the criterion for a successful transformation.

First of all, notice that (5.11) can be rewritten as

\[
\tilde{g}(x_2) - \tilde{g}(x_1) = \frac{g(x_2) - g(x_1)}{\sigma_1^2}, \quad (x_1 < x_2),
\]  

(8.1)
where

\[
\frac{1}{\sigma_{12}} = \frac{\int_{x_1}^{x_2} \frac{1}{\mu^{-1}(x)} \frac{1}{\sigma} g'(x) \, dy}{\int_{x_1}^{x_2} g'(x) \, dx}.
\]  \hspace{1cm} (8.2)

In an N'tf, \( g(X) \sim h(\nu, \sigma^2_\theta) \), this has the following interpretation: \( \tilde{g}(x_2) - \tilde{g}(x_1) \) is the number of standard deviations between \( g(x_1) \) and \( g(x_2) \), using the intermediate value \( \sigma_{12} \) as the unit of measurement. (Since \( g'(x) > 0 \), definition (8.2) necessarily gives \( \min \{ \sigma^{-1}_\mu(x) : x \in [x_1, x_2] \} < \sigma_{12} < \max \{ \sigma^{-1}_\mu(x) : x \in [x_1, x_2] \} \).) This is handy for the quick calculation of approximate significance levels, which is often the main point of making the transformation.

For the rest of this section we consider the family (4.15), \( X \sim h(\theta, (1+\epsilon \theta)^2) \) for \( \theta > -1/\epsilon, \epsilon \) a known constant. We have in mind values of \( \epsilon \) in the range \([0, 0.20]\). Here \( \mathcal{G} \) is already normal so \( g(x) = x \). The variance stabilizing transformation (5.5) is calculated to be

\[
\tilde{g}(x) = \frac{\log(1+\epsilon x)}{\epsilon}.
\]  \hspace{1cm} (8.3)

The transformed variate \( W = \tilde{g}(X) \) has a translation family of distributions

\[
W = \tilde{\theta} + \tilde{Z},
\]  \hspace{1cm} (8.4)

where

\[
\tilde{\theta} = \tilde{g}(\theta), \quad \tilde{Z} = \tilde{g}(Z),
\]  \hspace{1cm} (8.5)
\( Z \sim \mathcal{N}(0,1) \). Here we are ignoring the possibility \( 1+\epsilon z < 0 \), discussed in Section 4, an event with probability \( \Phi(-1/\epsilon) \leq \Phi(-5) \).

Suppose that we want a central \( 1-2\alpha \) confidence interval for \( \theta \) based on observing \( X = x \). The obvious \( 1-2\alpha \) central interval for \( \tilde{\theta} \) based on observing \( W = w \) in (8.4), is

\[
\tilde{\theta} \in \left[ w - \bar{g}(z_{1-\alpha}), w + \bar{g}(z_{\alpha}) \right]. \tag{8.5}
\]

This maps back to the \( 1-2\alpha \) central interval for \( \theta = \bar{g}^{-1}(\tilde{\theta}) \)

\[
\theta \in \left[ x + \hat{\sigma} \frac{z_{\alpha}}{1-\epsilon z_{\alpha}}, x + \hat{\sigma} \frac{z_{1-\alpha}}{1-\epsilon z_{1-\alpha}} \right], \tag{8.6}
\]

where \( \hat{\sigma} = 1+\epsilon x \). For reasonable values of \( \alpha \), say \( \alpha > .001 \), interval (8.6) corresponds to inverting the locally most powerful tests in family (4.15), and is close to being globally optimal, though it is not exactly so since this family does not enjoy monotone likelihood ratio. We refer to (8.6) as the true interval for \( \theta \) in what follows.

An easy approximate interval for \( \theta \) is

\[
\theta \in \left[ x + \hat{\sigma} z_{\alpha}, x + \hat{\sigma} z_{1-\alpha} \right]. \tag{8.7}
\]

This is obtained by first pretending that \( \sigma_\theta = 1+\epsilon \theta \) is a constant in (4.15), and then estimating the constant by \( \hat{\sigma} = 1+\epsilon x \).

Another approximate interval is obtained by ignoring the nonnormality of \( \tilde{Z} \) in (8.4). We suppose that

\[
\tilde{Z} \sim \mathcal{N}(0,\beta). \tag{8.8}
\]

Here we might take \( \beta = 1 \), since the transformation \( \tilde{g} \) is supposed to give unit variance, or use the actual variance of \( \frac{\log(1+\epsilon Z)}{\epsilon} \).
\[
\begin{align*}
\epsilon &= 0 & 0.05 & 0.1 & 0.2 \\
\beta &= 1 & 1.0063 & 1.0261 & 1.1231 \\
\end{align*}
\] (8.9)

In either case, assumptions (8.4), (8.8) lead to the interval

\[
\theta \in \left[ x + \hat{\sigma} \frac{e^{\beta z}}{\epsilon} \alpha, x^+ \frac{e^{\beta z}}{\epsilon} 1 - \alpha \right].
\] (8.10)

<table>
<thead>
<tr>
<th>\epsilon</th>
<th>True Interval (8.6)</th>
<th>Approximate (8.7)</th>
<th>Approximate (8.10), \beta = 1</th>
<th>Approximate (8.10), true \beta</th>
</tr>
</thead>
<tbody>
<tr>
<td>.05</td>
<td>[-1.520, 1.792] (1.656)</td>
<td>[-1.645, 1.645] (1.645)</td>
<td>[-1.579, 1.715] (1.647)</td>
<td>[-1.589, 1.726] (1.657)</td>
</tr>
<tr>
<td>.10</td>
<td>[-1.413, 1.969] (1.691)</td>
<td>[-1.645, 1.645] (1.645)</td>
<td>[-1.517, 1.788] (1.653)</td>
<td>[-1.553, 1.839] (1.696)</td>
</tr>
<tr>
<td>.20</td>
<td>[-1.238, 2.452] (1.845)</td>
<td>[-1.645, 1.645] (1.645)</td>
<td>[-1.402, 1.948] (1.675)</td>
<td>[-1.545, 2.235] (1.890)</td>
</tr>
</tbody>
</table>

Table 3. Comparison of the true intervals (8.6) with approximations (8.7) and (8.10), for \( \alpha = .05, x = 0, \hat{\sigma} = 1 \). Parenthetical numbers are 1/2 the interval length.

Table 3 compares the true interval (8.6) with (8.7) and with (8.10). The second approximation is seen to be better, more so if the correct value of \( \beta \) is used. In this highly simplified situation it is better to transform to homoscedasticity and ignore nonnormality than vice-versa. Of course one could always do a complete analysis and recover the true interval (8.6), working either with \( X \) or with \( W \). However, the practical motivation of transformation theory is to avoid complicated analysis,
especially in already complicated situations. One such situation is discussed next.

Suppose now that we observe independent variates \( X_i \sim N(\theta_i, (1-\varepsilon \theta_i)^2) \), \( i=1, 2, \ldots, n \). Corresponding to each observation is a \( 1 \times k \) vector of observed covariates \( m_i \). We intend to fit a linear model on either the \( X \) scale or the \( W \) scale (8.4). That is we will either fit the model

\[
\begin{align*}
\tilde{\theta} & = M\tilde{\alpha} \\
\sim & \\
\tilde{\theta} & = (\tilde{\theta}_1, \ldots, \tilde{\theta}_n)', M' = (m'_1, m'_2, \ldots, m'_n), \text{or the model}
\end{align*}
\]

\[
\bar{\theta} = (\bar{\theta}_1, \ldots, \bar{\theta}_n). \text{ The fitting will be done by ordinary least squares (OLS) in either case, to } \tilde{X} = \tilde{x} \text{ in (8.11) or to } \bar{W} = \bar{w} \text{ in (8.12). The question is, which of these analyses will be asymptotically most efficient for estimating the unknown } k \times 1 \text{ vector } \alpha, \text{ compared to maximum likelihood estimation?}
\]

Consider situation (8.12). The OLS estimate \( \tilde{\alpha} \) and the MLE \( \hat{\alpha} \) both have asymptotic covariance matrix of the form \( c(M'M)^{1/2} \). The ratio \( c_{\tilde{\alpha}}/c_{\alpha} \), which measures the asymptotic relative efficiency of \( \tilde{\alpha} \) to \( \hat{\alpha} \), say \( \text{Eff}_{\tilde{W}} \), turns out to be

\[
\text{Eff}_{\tilde{W}} = 1/\sqrt{(1+2\varepsilon^2)\beta}, \quad (8.13)
\]

\( \beta \) as given in (8.9), see Cox and Hinkley (1968).

Situation (8.11) is less neat, since in this case the OLS estimate \( \tilde{\alpha} \) does not have a covariance matrix of form \( c(M'M)^{1/2} \). As a measure of efficiency comparable to (8.13) we take
\begin{equation}
\text{Eff}_X = \left| \frac{\hat{\Sigma}_\alpha}{\hat{\Sigma}_\alpha \text{d}^{-1}} \right|^{1/(2k)} \left| \frac{\hat{\Sigma}_\alpha}{\hat{\Sigma}_\alpha \text{d}^{-1}} \right|^{1/(2k)}, \tag{8.14}
\end{equation}

where $\hat{\Sigma}_\alpha$ and $\hat{\Sigma}_\alpha \text{d}^{-1}$ are the asymptotic covariance matrices. Using results of Bloomfield and Watson (1975), one can show that

\begin{equation}
\text{Eff}_X \geq \frac{2\sqrt{\text{RATIO}}}{1 + \text{RATIO}}, \tag{8.15}
\end{equation}

\[ \text{RATIO} \equiv \left[ \frac{1 + \varepsilon(\max \theta_i)}{1 + \varepsilon(\min \theta_i)} \right]^2. \]

The lower bound (8.15) on $\text{Eff}_X$ is achieved if the design matrix $\tilde{M}$ has a certain relationship to the covariance matrix of $X$.

\begin{table}[h]
\centering
\begin{tabular}{c|cccc}
\multicolumn{1}{c|}{} & 0 & .05 & .1 & .20 \\
\hline
\text{Eff}_W & 1 & .9944 & .9775 & .9080 \\
\text{Eff}_X & & 1 & .9975 & .9901 & .9623 \\
\text{Lower Bound} & 1 & .9798 & .9774 & .9701 & .9428 \\
\text{RATIO} & 1.5 & .9428 & .9405 & .9335 & .9072 \\
\text{Eff}_X & 2 & .8660 & .8638 & .8575 & .8333 \\
\text{Eff}_X & 3 & & & & \\
\end{tabular}
\caption{Comparison of the asymptotic efficiency of ordinary least squares on the variance stabilized scale, (8.13), with the lower bound for efficiency on the normalized scale, (8.15).}
\end{table}

Table 4 compares $\text{Eff}_W$ with the lower bound for $\text{Eff}_X$. If $\text{RATIO} = 1$ then efficiency is always better on the $X$ scale, but for larger values of
RATIO, which are probably more realistic, the W scale seems preferable. For moderate values of $\varepsilon$, $|\varepsilon| \leq .1$, efficiency on the W scale can't be much worse than 98%, which is quite safe indeed.
References


