The Annals of Statistics
1982, Vol. 10, No. 2, 323–339

THE 1981 WALD MEMORIAL LECTURES

TRANSFORMATION THEORY: HOW NORMAL IS A FAMILY OF DISTRIBUTIONS?[1]

BY BRADLEY EFRON

Stanford University

This paper concerns the following question: if X is a real-valued random variate having a one-parameter family of distributions $\mathscr{F}$, to what extent can $\mathscr{F}$ 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 $\mathscr{F}$? The theory developed answers the question without considering the form of g at all. In those cases where the answer is positive, simple formulas for calculating g are given. The paper also considers the relationship between normalization and variance stabilization.

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

$$Y = \tanh^{-1} X = \frac{1}{2} \ln\left(\frac{1+X}{1-X}\right) \tag{1.1}$$

has, approximately, a normal distribution

$$Y \sim N\left(\nu_\theta, \frac{1}{n-3}\right), \tag{1.2}$$

where

$$\nu_\theta = \tanh^{-1}\theta + \frac{\theta}{2(n-1)}. \tag{1.3}$$

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? Firstly, because quick calculations of significance levels are much easier on the Y scale. For example, with $n = 15$, $x = 0.70$ is significantly different from $\theta = 0.20$, because $\sqrt{n-3}\,(\mathrm{y} - \nu_{0.2}) = 2.277$, a significant normal deviate at level .011, one-sided. Secondly, normal theory methods can be applied on the Y scale. Suppose $X_1 = x_1, X_2 = x_2, \cdots, X_J = x_J$ are observed sample correlations calculated from independent data sets, and we are interested in the possible relationship of these correlations to covariate vectors $c_1, c_2, c_3, \cdots, c_J$. A standard regression analysis of the transformed values $y_1, y_2, \cdots, y_J$ versus $c_1, c_2, \cdots, c_J$ is the natural way to proceed. Thirdly, appropriate confidence intervals are easy to calculate on the Y scale. For example, the 90% central interval

$$\nu_\theta \in y \pm 1.645/\sqrt{n-3}$$

gives the usual interval for θ, by inverting function (1.3).

Received August 1981; revised December 1981.

[1] This research supported by Public Health Service Grant 2 R01 GM21215-07 and National Science Foundation Grant MCS80-24649.

AMS 1980 *subject classifications*: 62E10, 62E99.

Key words and phrases. Normalization, variance stabilization, square root transformation, power transformations.

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 $\mathscr{F}$, the correlation distributions, to a simpler family $\mathscr{G}$, the normal translation family. The appeal of (1.1)–(1.3) is similar to representing a symmetric matrix F as $\Gamma G \Gamma'$, where Γ 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 $\mathscr{F}$ and $\mathscr{G}$.

This paper concerns the following question: if X is a real-valued random variate having a one-parameter family of distributions $\mathscr{F}$, to what degree can $\mathscr{F}$ 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 $\mathscr{F}$?

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

$$F_\theta(x) = \Phi\left(\frac{g(x) - \nu_\theta}{\sigma_\theta}\right). \tag{1.4}$$

Here Φ is the standard normal cdf, θ is the real-valued parameter indexing $\mathscr{F}$, ν_θ is the median of $g(X)$, and σ_θ the standard deviation of $g(X)$.

For two different values of θ, say θ_1 and θ_2, define

$$z_i(x) \equiv \Phi^{-1}(F_{\theta_i}(x)) = \{g(x) - \nu_{\theta_i}\}/\sigma_{\theta_i}, \tag{1.5}$$

$i = 1, 2$, the last equality following from (1.4). Eliminating $g(x)$ from the two equations (1.5) gives

$$z_2(x) = \frac{\sigma_{\theta_1}}{\sigma_{\theta_2}} z_1(x) + \frac{\nu_{\theta_1} - \nu_{\theta_2}}{\sigma_{\theta_2}}. \tag{1.6}$$

The quantities $z_i(x)$ can be calculated directly from the cdf's $F_{\theta_i}(x)$, without any knowledge of g. Equation (1.6) shows that *if $\mathscr{F}$ 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 (1.6) we see that this implies $\sigma_{\theta_2}/\sigma_{\theta_1} \doteq 1$, in close agreement with (1.2).

The diagnostic function $D(z, \theta)$ introduced in Section 2 is a more convenient way of

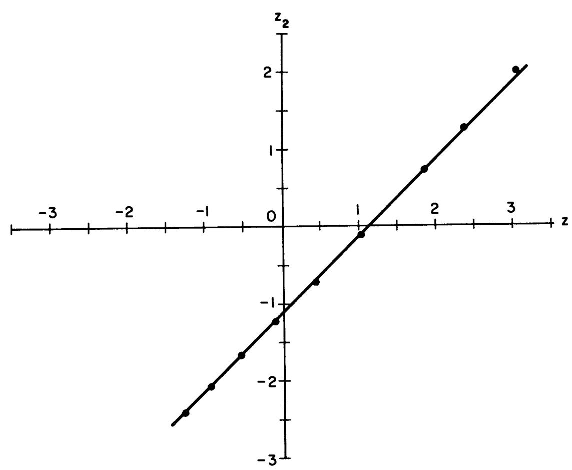

FIG. 1. *A plot of $z_2(x)$ versus $z_1(x)$ for the normal correlation family, $n = 15$, $\theta_1 = 0.5$, $\theta_2 = 0.7$. The plot is nearly linear, indicating the existence of a nearly normalizing transformation.*

carrying out the same analysis, without having to consider all pairs of values θ_1 and θ_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).

Fisher's transformation for the correlation coefficient is *stabilizing* as well as *normalizing*. The function "σ_θ" is constant in (1.2). The $\tanh^{-1}$ transformation produces nearly stable variances as well as normality. For other families $\mathscr{F}$ there is tension between the twin goals of stabilization and normalization. For example, in the Poisson family, $\sqrt{(x + 3/8)}$ is an excellent variance stabilizer, while $x^{2/3}$ is an excellent normalizer; see Anscombe (1948, 1953).

The second purpose of this paper is to examine the relationship between stabilization and normalization. For instance 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 $\mathscr{F}$, 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 $\mathscr{F}$, a one-parameter family of distributions for the real-valued continuous variate X. Let

$$F_\theta(x) = \mathrm{Prob}_\theta\{X \le x\}$$

be the cumulative distribution function of X for parameter value θ, where $\theta \in \Theta$ the parameter space, a possibly infinite interval of the real line. The derivatives $\dot{F}_\theta(x) = (\partial/\partial\theta)\, F_\theta(x)$ and $f_\theta(x) = (\partial/\partial x)\, F_\theta(x)$ are assumed to exist in what follows. We wonder whether $\mathscr{F}$ is a *normal transformation family*, abbreviated NTF, that is whether there exists a strictly monotonic transformation $g(x)$ such that

$$g(X) \sim N(\nu_\theta, 1) \tag{2.1}$$

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

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

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

so that $x_{\alpha,\theta}$ is the 100 αth percentile point for X under F_θ. In particular, $x_{.5,\theta}$ is the median of X. Then the diagnostic function is defined as

$$D(z, \theta) = \frac{\dot{F}_\theta(x_{\Phi(z),\theta})}{\dot{F}(x_{.5,\theta})} \frac{\phi(0)}{\phi(z)}, \tag{2.2}$$

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 in terms of the cdf's $F_\theta(x)$ above, so that it can be evaluated without knowledge of $g(x)$, or even the assumption that $g(x)$ exists.

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

$$t_\theta(x) \equiv \Phi^{-1}F_\theta(x). \tag{2.3}$$

Under parameter value θ, $t_\theta(X)$ has a $N(0, 1)$ distribution. Definition (2.2) can be rewritten as

$$D(z, \theta) = \frac{\dot{t}_\theta(x_{\Phi(z),\theta})}{\dot{t}_\theta(x_{.5,\theta})}, \tag{2.4}$$

where $\dot{t}_\theta(x) \equiv (\partial/\partial\theta)\, t_\theta(x)$.

Without going into details, $D(z, \theta)$ measures how quickly the local transformation to normality is changing as θ varies. (In a NTF family (2.1), $t_\theta(x) = g(x) - \nu_\theta$; any local transformation to normality globally normalizes $\mathscr{F}$ in this case; see Section 5.)

It turns out that $D(z, \theta) \equiv 1$ if $\mathscr{F}$ is a normal transformation family. More usefully, plots of $D(z, \theta)$ enables us to diagnose deviations of $\mathscr{F}$ from the ideal form (2.1). To this end, consider a more general family $\mathscr{F}$ satisfying

$$g(X) \sim \nu_\theta + \sigma_\theta q(Z) \tag{2.5}$$

for some strictly monotonic transformation $g(x)$. Here and throughout, Z denotes a standard normal deviate; $q(Z)$ is a strictly increasing differentiable function satisfying

$$q(0) = 0, \qquad q'(0) = 1; \tag{2.6}$$

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

For a normal transformation family (2.1), $\sigma_\theta \equiv 1$ and $q(z) \doteq z$. The form (2.5) allows the scaling parameter σ_θ to vary with θ, 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 $\mathscr{F}$ satisfying (2.5) for some choice of $g(x)$, $q(z)$, ν_θ, and σ_θ a *general scaled transformation family*, abbreviated GSTF.

LEMMA 1. *If $\mathscr{F}$ is a general scaled transformation family, then the diagnostic function equals*

$$D(z, \theta) = \frac{1 + q(z)\varepsilon_\theta}{q'(z)}, \tag{2.7}$$

where

$$\varepsilon_\theta = \frac{\dot{\sigma}_\theta}{\dot{\nu}_\theta} = \frac{\partial\sigma_\theta/\partial\theta}{\partial\nu_\theta/\partial\theta}.$$

(Proof later in this section.)

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

Suppose $q(z) = z$ but that σ_θ is not a constant, a situation called a *normal scaled transformation family*, NSTF, in Section 3. Then the lemma gives

$$D(z, \theta) = 1 + z\varepsilon_\theta. \tag{2.8}$$

As an example, consider the continuous Poisson family of cdf's defined by

$$F_\theta(x) \equiv \frac{\int_\theta^\infty t^{x-1/2}e^{-t}\,dt}{\Gamma\left(x + \frac{1}{2}\right)}, \qquad x > -\frac{1}{2}, \tag{2.9}$$

$\Theta = (0, \infty)$. This family relates to the usual discrete Poisson family as follows: if x_0 is a nonnegative integer then

$$F_\theta(x_0 + ½) = \text{Prob}\{G_{x_0+1} > \theta\} = \text{Prob}\{Po(\theta) < x_0\}, \tag{2.10}$$

where G_{x_0+1} is a gamma variate with shape parameter $x_0 + 1$ and $Po(\theta)$ is a standard Poisson variate with parameter θ. (The last equality in (2.10) is well-known from the theory of Poisson processes.) In other words, the cdf of the continuous Poisson with parameter θ agrees with the cdf of the standard Poisson distribution, mean θ, 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.

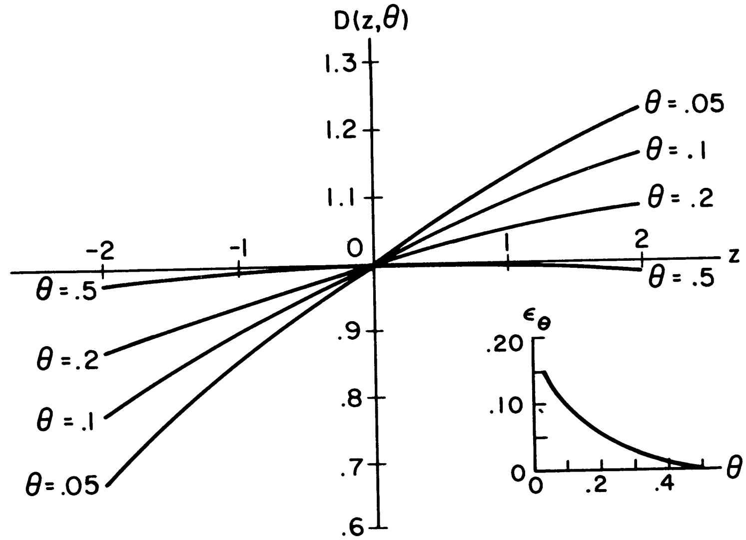

FIG. 2. *The diagnostic function* $D(z, \theta)$ *for the continuous Poisson family. Insert shows* $\varepsilon_\theta = \dot{\sigma}_\theta/\dot{\nu}_\theta$.

Figure 2 illustrates $D(z, \theta)$ for the continuous Poisson family. Good agreement with (2.8) is evident, even for small values of θ. The function $\varepsilon_\theta = \dot{\sigma}_\theta/\dot{\nu}_\theta$ 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 σ_θ increasing rather rapidly as a function of the median ν_θ, especially for small values of θ. These points are discussed further in Sections 4 and 6. Section 7 gives a more complete discussion of the Poisson family.

Lemma 1 enables us to calculate $q(z)$ from $D(z, \theta)$ after which we can also calculate the functions $g(x)$, ν_θ, and σ_θ; see Sections 4, 5, and 6. This means that the GSTF representation (2.5) of a family $\mathscr{F}$ is unique, with one interesting exception discussed in Section 4. The 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 Lemma 1. The cdf of $\tilde{Z} = q(Z)$ is

$$\tilde{\Phi}(\tilde{z}) = \Phi(q^{-1}(\tilde{z})). \tag{2.11}$$

If z_α is the 100αth normal percentile,

$$\Phi(z_\alpha) = \alpha,$$

then $\tilde{z}_\alpha = q(z_\alpha)$ is the corresponding percentile of $\tilde{Z}$. In particular, $\tilde{Z}$ has median $q(z_{.5}) = q(0) = 0$, by (2.6). The density function of $\tilde{Z}, \tilde{\phi}(\tilde{z}) = \tilde{\phi}'(\tilde{z})$, satisfies

$$\tilde{\phi}(\tilde{z}) = \phi(z)/q'(z), \tag{2.12}$$

so

$$\tilde{\phi}(0) = \phi(0) = \frac{1}{\sqrt{2\pi}} \tag{2.13}$$

by (2.6).

From (2.5) we get $F_\theta(x) = \tilde{\Phi}\left(\dfrac{g(x) - \nu_\theta}{\sigma_\theta}\right)$, implying that

$$\dot{F}_\theta(x) = -\tilde{\phi}\left(\frac{g(x) - \nu_\theta}{\sigma_\theta}\right)\left\{\frac{\dot{\nu}_\theta}{\sigma_\theta} + \frac{g(x) - \nu_\theta}{\sigma_\theta}\frac{\dot{\sigma}_\theta}{\sigma_\theta}\right\}.$$

But, again using (2.5) and the fact that percentiles map in the obvious way under a monotone transformation, that is

$$\tilde{z}_\alpha = \frac{g(x_{\alpha,\theta}) - \nu_\theta}{\sigma_\theta},$$

we have

$$\dot{F}_\theta(x_{\alpha,\theta}) = -\tilde{\phi}(\tilde{z}_\alpha)\left(\frac{\dot{\nu}_\theta}{\sigma_\theta} + \tilde{z}_\alpha \frac{\dot{\sigma}_\theta}{\sigma_\theta}\right) = -\frac{\phi(z_\alpha)}{q'(z_\alpha)}\left\{\frac{\dot{\nu}_\theta}{\sigma_\theta} + q(z_\alpha)\frac{\dot{\sigma}_\theta}{\sigma_\theta}\right\}$$

by (2.12). In particular $\dot{F}_\theta(x_{.5,\theta}) = -\phi(0)(\dot{\nu}_\theta/\sigma_\theta)$. Substituting these expressions into (2.4) gives (2.7). □

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

The proof of these statements is based on (2.4). First of all, we can assume that

$$\dot{t}_\theta(x_{.5,\theta}) = -1. \tag{2.14}$$

If not, then a change of parameters makes it so: define the new parameter ξ by

$$\frac{d\xi}{d\theta} = \frac{f_\theta(\mu_\theta)}{\phi(0)}\dot{\mu}_\theta, \tag{2.15}$$

where $\mu_\theta = x_{.5,\theta}$, and $f_\theta(x) = (\partial/\partial x)F_\theta(x)$. For convenience we assume $\dot{\mu}_\theta > 0$. It is easy to verify that $(\partial/\partial\xi)t_\theta(x)|_{x=\mu_\theta} = -1$. Then (2.14) holds with $\mathscr{F}$ parameterized by ξ instead of θ. Reparameterizations have no effect on the D function, as commented earlier, nor on the interpretations of D based on Lemma 1. The ξ parameterization (2.16) is unique (up to an additive constant) for any family $\mathscr{F}$. Here we assume $\theta = \xi$, for ease of notation.

Choose any set of x values, say $x_1, x_2, \cdots, x_K$, and define $\mathbf{z}_\theta = (t_\theta(x_1), t_\theta(x_2), \cdots, t_\theta(x_K))$. From (2.4) and (2.14) it follows that the derivative of $\mathbf{z}_\theta$ with respect to θ is

$$\dot{\mathbf{z}}_\theta = -(D(z_{\theta 1}, \theta), D(z_{\theta 2}, \theta), \cdots, D(z_{\theta K}, \theta)), \tag{2.16}$$

$z_{\theta k} \equiv t_\theta(x_k)$. As θ moves through Θ, the vector $\mathbf{z}_\theta$ traces out a curve in $\mathscr{R}^K$ completely describing how the values of $F_\theta(x_k)$, $k = 1, \cdots, K$ relate to one another. This curve is determined by the differential equation (2.16), which depends on the function $D(z, \theta)$.

Suppose that $D(z, \theta) = 1 + z\varepsilon_\theta$ as at (2.9). It now follows that $\mathscr{F}$ must be an NSTF, without first assuming as we previously did that $\mathscr{F}$ is a GSTF. If $D(z, \theta)$ is "nearly" of the form $1 + z\varepsilon_\theta$ then $\mathscr{F}$ must "nearly" be an NSTF, in a sense made precise by numerically solving (2.16) 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 $\mathscr{F}$ agrees with the normal transformation form (2.1). To this end it is useful to define more general types of transformation family representing various departures from (2.1). Two such generalizations have already been introduced: the general scaled transformation family (2.5), and the normal scaled transformation family referred to at (2.8).

Table 1 describes the six types of transformation family used in this paper. The most general case, GSTF, represents a given family $\mathscr{F} = \{X \sim F_\theta, \theta \in \Theta\}$ in terms of a standard normal variate Z as follows: Z is transformed to $\tilde{Z} = q(Z)$ by a strictly increasing mapping $q(z)$; $Y = \nu_\theta + \sigma_\theta \tilde{Z}$ 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

TABLE 1

Six transformation types, described in terms of a standard normal variate Z and the four functions g^{-1}, ν_θ, σ_θ, 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 θ_0.

Description	Name and Abbreviation	Relationship
1. $X = g^{-1}(\nu_\theta + Z)$	Normal transformation family (NTF)	NTF
2. $X = g^{-1}(\nu_\theta + q(Z))$, $q(-z) = -q(z)$	Symmetric transformation family (STF)	NSTF
3. $X = g^{-1}(\nu_\theta + q(Z))$	General transformation family (GTF)	STF
4. $X = g^{-1}(\nu_\theta + \sigma_\theta Z)$	Normal scaled transformation family (NSTF)	SSTF
5. $X = g^{-1}(\nu_\theta + \sigma_\theta q(Z))$, $q(-z) = -q(z)$	Symmetric scaled transformation family (SSTF)	GTF
6. $X = g^{-1}(\nu_\theta + \sigma_\theta q(Z))$	General scaled transformation family (GSTF)	GSTF

with g as in Section 2.) In addition to restrictions (2.6) on $q(z)$, we set

$$\nu_{\theta_0} = 0, \qquad \sigma_{\theta_0} = 1 \tag{3.1}$$

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 through 5 in Table 1 represent intermediates between the simple NTF case (2.1) and the GSTF case (2.5). In type 5, for example, SSTF, $\tilde{Z} = q(Z)$ is restricted to be symmetrically distributed about 0. Section 5 shows that some calculations are easier in the SSTF case than in a GSTF.

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

The upper set of curves applies to $n = 15$ bivariate normal points. We shall see (Section 4) 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. This is as close as we come in this paper to a genuine example of a normal transformation family.

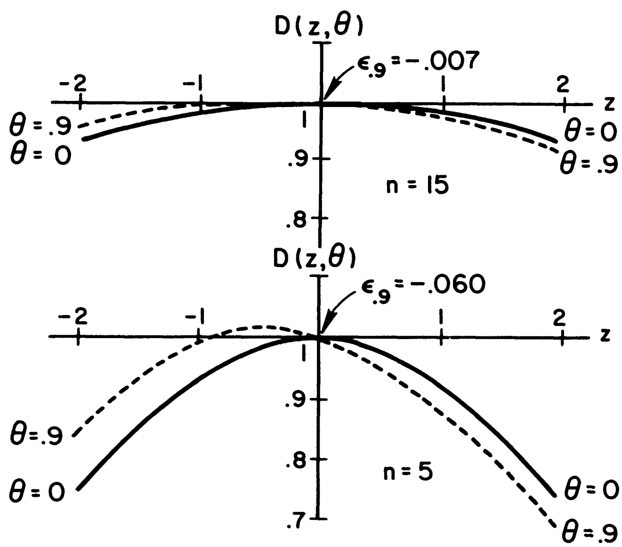

FIG. 3. *The diagnostic function $D(z, \theta)$ for the normal correlation coefficient, $\theta = \rho$, $X = \hat{\rho}$. The upper curves are for sample size $n = 15$, the lower for $n = 5$.*

The lower set of curves applies to $n = 5$. This turns out to be close to an SSTF situation, with $\tilde{Z} = q(Z)$ a Student's t variate with 10 degrees of freedom. The scale function σ_θ has its maximum at $\theta = \rho = 0$, and decreases as θ goes to ± 1: at $\theta = .5$, $\varepsilon_\theta = \dot{\sigma}_\theta / \dot{\nu}_\theta$ equals $-.041$; at $\theta = .9$ it equals $-.060$.

If we assume that the normal correlation family $\mathscr{F}$ is a GSTF, then it must be a SSTF. This follows from the symmetry of $\mathscr{F}$ about 0, i.e. the fact that the mapping $(\theta, X) \to (-\theta, -X)$ takes $\mathscr{F}$ into itself, and the uniqueness of the GSTF 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 $1 - 2\alpha$ central confidence interval for θ based on observing $Y = g(X)$,

$$\nu_\theta \in [y - \tilde{z}_{1-\alpha}, y - \tilde{z}_\alpha], \tag{3.2}$$

where $\tilde{z}_\alpha$ is the 100 αth percentile of $\tilde{Z} = q(Z)$. This transforms back into a confidence interval for θ, assuming we know the mapping from ν_θ back to θ. 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, $\sigma_\theta \equiv 1$, so $\varepsilon_\theta = \dot{\sigma}_\theta / \dot{\nu}_\theta = 0$ and Lemma 1 gives

$$D(z, \theta) = \frac{1}{q'(z)}$$

not depending upon θ. It will turn out (Section 4) that $D(z, \theta)$ not depending upon θ is a sufficient as well as necessary condition for the simple types NTF, STF, GTF.

4. Finding $q(Z)$. The representation $X = g^{-1}(\nu_\theta + \sigma_\theta q(Z))$ involves four functions: g, ν_θ, σ_θ, 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 (2.2). For convenient discussion, henceforth "GSTF" 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 $\mathscr{F}$ is a* GSTF, $X = g^{-1}(\nu_\theta + \sigma_\theta q(Z))$, *and that there exist two values θ_1 and θ_2 such that $D(z, \theta_1)$ is not identically equal to $D(z, \theta_2)$. Then*

$$\frac{d}{dz} \log\{q(z)/z\} = \frac{D'(0, \theta_2) - D'(0, \theta_1)}{D(z, \theta_2) - D(z, \theta_1)} - \frac{1}{z}, \tag{4.1}$$

where $D'(0, \theta) = \partial/\partial z D(z, \theta)|_{z=0}$. If $D(z, \theta)$ does not depend on θ, then $\mathscr{F}$ is a GTF, $X = g^{-1}(\nu_\theta + q(Z))$, *and*

$$q'(z) = \frac{1}{D(z)}. \tag{4.2}$$

PROOF. From Lemma 1 and (2.7) we get

$$D(z, \theta_2) - D(z, \theta_1) = (\varepsilon_{\theta_2} - \varepsilon_{\theta_1}) \frac{q(z)}{q'(z)} \tag{4.3}$$

for any $\theta_1, \theta_2 \in \Theta$. Also, using (2.6),

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

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

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

which is (4.1).

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

$$\sigma_\theta = 1 + c\nu_\theta$$

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

$$1 + cg(X) = (1 + c\nu_\theta) + (1 + c\nu_\theta)c\tilde{Z} = (1 + c\nu_\theta)(1 + c\tilde{Z}). \tag{4.5}$$

Letting $g^\circ(x) = c^{-1} \log\{1 + cg(x)\}$, $\nu_\theta^\circ = c^{-1} \log(1 + c\nu_\theta)$, and $\tilde{Z}^\circ = c^{-1} \log(1 + c\tilde{Z})$ gives

$$g^\circ(X) = \nu_\theta^\circ + \tilde{Z}^\circ, \tag{4.6}$$

a GTF representation. (Notice that $q^\circ(Z) = c^{-1} \log(1 + c\tilde{Z}) = c^{-1} \log\{1 + cq(Z)\}$ satisfies (2.6). The difficulty with $1 + c\tilde{Z}$ possibly being negative is discussed below.) But in a GTF 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 GSTF case, the theorem allows us to calculate $d \log\{q(z)/z\}/dz$ 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 $\lim_{z\to 0} q(z)/z = 1$, 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, ν_θ, and σ_θ 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

$$g^\circ(X) = \nu_\theta^\circ + (1 + c\nu_\theta^\circ)q^\circ(Z), \tag{4.7}$$

where c is any constant and $g^\circ(x) = [\exp\{cg(x)\} - 1]/c$, $\nu_\theta^\circ = \{\exp(c\nu_\theta) - 1\}/c$, and $q^\circ(z) = [\exp\{cq(z)\} - 1]/c$. Representation (4.7) is a GSTF with $\varepsilon_\theta = \dot{\sigma}_\theta/\dot{\nu}_\theta = c$. In other words, corresponding to any GTF is a one-parameter family of GSTF representations, with the free parameter being the constant value of ε_θ. The GTF representation, having $\varepsilon_\theta \equiv 0$, is obtained from (4.2). There is only one such representation for a given GTF family $\mathscr{F}$, 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 SSTF situation so that $q(-z) = -q(z)$. Writing

$$q(z) = z + \frac{Bz^3}{6} + \cdots, \tag{4.8}$$

(2.7) gives

$$D(z, \theta) = 1 + \varepsilon_\theta z - \frac{B}{3} z^2 \cdots. \tag{4.9}$$

Stopping the series after the quadratic term gives a reasonable approximation to $D(z, \theta)$ for the normal correlation coefficient, Figure 3. For $n = 5$, $B \doteq .146$; for $n = 15$, $B \doteq .039$.

The Cornish-Fisher expansion for a Student-t variate with d degrees of freedom, rescaled to have the same density as a $N(0, 1)$ at $z = 0$, begins $q(z) = z + z^3/(4d + 2)$; see Johnson and Kotz (1970, page 102). Comparing this with (4.8) gives the approximation

$$d \doteq \frac{1}{4}\left(\frac{6}{B} - 2\right). \tag{4.10}$$

This gives $d \doteq 10$ for $n = 5$ and $d \doteq 38$ 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 (4.8), (4.9). 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 $N(0, 1)$ variate, so the interpretation of the continuous Poisson family as an NSTF seems quite reasonable.

In an SSTF, we have the simple relationship

$$\varepsilon_\theta = D'(0, \theta), \tag{4.11}$$

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

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

In an SSTF, $q''(0) = 0$ by symmetry, giving (4.11).

In going from (4.5) to (4.6) we ignored the possibility $1 + c\tilde{Z} < 0$. The following special case illustrates what happens in this situation. Let $\mathscr{F}$ be the family

$$X \sim N(\theta, (1 + \varepsilon\theta)^2), \qquad \theta > -1/\varepsilon. \tag{4.12}$$

Here ε is a positive constant. This family is an NSTF $q(z) = z$, with g the identity mapping, $\nu_\theta = \theta$, $\sigma_\theta = 1 + \varepsilon\theta$. By (2.8), $D(z, \theta) = 1 + \varepsilon z$. We can also write (4.12) as

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

$Z \sim N(0, 1)$. The two sign cases can be separately transformed,

$$\begin{aligned} \frac{\log(1 + \varepsilon X)}{\varepsilon} &= \frac{\log(1 + \varepsilon\theta)}{\varepsilon} + \frac{\log(1 + \varepsilon Z)}{\varepsilon} \quad \text{for} \quad 1 + \varepsilon X > 0 \\ \frac{\log-(1 + \varepsilon X)}{\varepsilon} &= \frac{\log(1 + \varepsilon\theta)}{\varepsilon} + \frac{\log-(1 + \varepsilon Z)}{\varepsilon} \quad \text{for} \quad 1 + \varepsilon X < 0. \end{aligned} \tag{4.13}$$

In other words, (4.12) is a GTF "on two real lines," one real line corresponding to each sign of $1 + \varepsilon X$. The parameter $\nu_\theta^\circ = (1/\varepsilon)\log(1 + \varepsilon\theta)$ translates the distribution of $X^\circ = (1/\varepsilon)\log|1 + \varepsilon X|$ 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 ν_θ° varies. Formula (4.2),

$$q'(z) = \frac{1}{1 + \varepsilon z}$$

in this case, gives both transformations of Z in (4.13).

To summarize, if $D(z, \theta)$ does not depend on θ then $\mathscr{F}$ is a GTF, though possibly defined on two lines. Formulas (4.1) and (4.2) give 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

$$\frac{d}{dz}\log|q(z)| = \frac{\dfrac{\partial}{\partial\theta}D'(0, \theta)}{\dfrac{\partial}{\partial\theta}D(z, \theta)}. \tag{4.14}$$

In a GSTF (4.1), or (4.14), does not depend on the choice of θ values. If this is markedly untrue then $\mathscr{F}$ does not have a good GSTF approximation.

It is easy to see when a GSTF family $\mathscr{F}$ is actually SSTF. The function $D_+(z, \theta) \equiv D(z, \theta) + D(-z, \theta)$ equals

$$D_+(z, \theta) = \frac{2}{q'(z)} + \varepsilon_\theta \frac{q(z) + q(-z)}{q'(z)}$$

in a GSTF. Assuming that ε_θ is not constant, i.e. that $\mathscr{F}$ is not a GTF family, then $D_+(z, \theta)$

not depending on θ is necessary and sufficient for $\mathscr{F}$ to be SSTF, since both conditions are equivalent to $q(z) + q(-z) = 0$. If $\mathscr{F}$ is a GTF then (4.2) gives $D(z) = D(-z)$ as necessary and sufficient for $\mathscr{F}$ to be STF.

5. Finding $g(X)$. We wish to compute the function g in the GSTF representation $g(X) = \nu_\theta + \sigma_\theta q(Z)$. Two formulas will be given, one for the general GSTF case and a simpler one applying to the GTF case. Let $x_1 < x_2$ be any two values of X, and define θ_{12} as that value of θ making $F_\theta(x_1) = 1 - F_\theta(x_2)$, say

$$\alpha = F_{\theta_{12}}(x_1) = 1 - F_{\theta_{12}}(x_2). \tag{5.1}$$

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

$$F_{\theta_x}(x) = .5,$$

and let

$$f_\theta(x) = F'_\theta(x),$$

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

THEOREM 2. *In a* GSTF *family, under definition* (5.1),

$$\frac{g'(x_2)}{g'(x_1)} = \frac{f_{\theta_{12}}(x_2)}{f_{\theta_{12}}(x_1)} \frac{q'(z_{1-\alpha})}{q'(z_\alpha)}. \tag{5.2}$$

The simpler formula

$$\bar{g}'(x) = \frac{f_{\theta_x}(x)}{\phi(0)} \tag{5.3}$$

gives $g'(x)$ in a GTF *family.*

PROOF. Since $F_\theta(x) = \tilde{\Phi}\left(\frac{g(x) - \nu_\theta}{\sigma_\theta}\right)$ where $\tilde{\Phi}(\tilde{z}) = \Phi(q^{-1}(\tilde{z}))$ as at (2.11), differentiation yields $f_\theta(x) = \tilde{\phi}\left(\frac{g(x) - \nu_\theta}{\sigma_\theta}\right) \frac{g'(x)}{\sigma_\theta}$. Substituting $x = x_{\alpha,\theta}$ gives

$$f_\theta(x_{\alpha,\theta}) = \tilde{\phi}(\tilde{z}_\alpha) \frac{g'(x_{\alpha,\theta})}{\sigma_\theta} = \frac{\phi(z_\alpha) g'(x_{\alpha,\theta})}{q'(z_\alpha)\sigma_\theta}, \tag{5.4}$$

the last equality following from (2.12). But for $\theta = \theta_{12}$, by (5.1), we have $x_1 = x_{\alpha,\theta}$, $x_2 = x_{1-\alpha,\theta}$, and $\phi(z_\alpha) = \phi(z_{1-\alpha})$, so (5.4) follows from (5.2) by division.

The GTF formula (5.3) follows from the last equality in (5.4). We take $\theta = \mu^{-1}(x) = \theta_x$, $\alpha = .5$, so $x = x_{\alpha,\theta}$ and $z_\alpha = 0$. Since $\sigma_\theta = 1$ in a GTF, and $q'(0) = 1$ by (2.6), (5.4) gives $g'(x) = f_{\theta_x}(x)/\phi(0)$ as claimed. □

Formula (5.2) simplifies in the SSTF situation to

$$\frac{g'(x_2)}{g'(x_1)} = \frac{f_{\theta_{12}}(x_2)}{f_{\theta_{12}}(x_1)}, \tag{5.5}$$

since $q'(z_{1-\alpha}) = q'(z_\alpha)$ by symmetry. Like (5.3), formula (5.5) has the advantage of not requiring knowledge of $q(z)$.

Formula (5.2), and (5.5) in the SSTF 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 α fixed, we calculate $x_2, x_3, x_4, \cdots$ and $\theta_{12}, \theta_{23}, \theta_{34}, \cdots$ satisfying $\alpha = F_{\theta_{i-1,i}}(x_{i-1}) = 1 - F_{\theta_{i-1,i}}(x_i)$. Successive use of (5.2) gives $g'(x_1), g'(x_2), g'(x_3), \cdots$ 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 (6.3).

Letting $\alpha \to .5$ in (5.2) gives a "single x" version,

$$\frac{d}{dx}\log g'(x) = \frac{d}{dx}\log f_\theta(x)\,|_{\theta=\theta_x} + q''(0)f_{\theta_x}(x)/\phi(0).$$

The last term vanishes in an SSTF, since $q''(0) = 0$.

Formula (5.3) has a simple intuitive interpretation in terms of the local transformation to normality $t_\theta(x) = \Phi^{-1}F_\theta(x)$, (2.3). In an NTF, $X = g^{-1}(\nu_\theta + Z)$, we have $t_{\theta_0}(x) = \Phi^{-1}\Phi(g(x) - \nu_{\theta_0}) = g(x) - \nu_{\theta_0}$, so

$$t'_{\theta_0}(x) = g'(x). \tag{5.6}$$

This means that it doesn't matter which value of θ_0 we choose: $t'_{\theta_0}(x)$ always agrees with $g'(x)$ in an NTF. In other words, any local transformation to normality globally normalizes an NTF.

If $\mathscr{F}$ is not an NTF then (5.6) doesn't hold. However, we can try to choose among the different $t_\theta(x)$ transformations by selecting that θ most appropriate to each x. An obvious choice is $\theta = \theta_x = \mu^{-1}(x)$, with $t_\theta(x)$ having x derivative $t'_\theta(x)$ equaling

$$t'_{\theta_x}(x) = f_{\theta_x}(x)/\phi(0),$$

which is formula (5.3). In words, $\bar{g}(x)$ is the transformation everywhere having the same x derivative as $t_\theta(x)$, evaluated at that θ for which x is the median of X.

In an important sense $\bar{g}(x)$ deserves to be called a *variance stabilizing transformation*. In a GTF, where perfect stabilization is possible, $\bar{g}(x)$ achieves this exactly: $Y = \bar{g}(X) = \nu_\theta + \tilde{Z}$, a translation family, with constant variance.

The following corollary shows that $\bar{g}(x)$ tries to stabilize variances in the more general context of a GTF.

COROLLARY 1. *If $\mathscr{F}$ is a* GSTF, $X \sim g^{-1}(\nu_\theta + \sigma_\theta\tilde{Z})$, *then*

$$\bar{g}'(x) = g'(x)/\sigma_{\theta_x}. \tag{5.7}$$

PROOF. Taking $\alpha = .5$ in (5.4) gives

$$f_{\theta_x}(x) = \frac{g'(x)}{\sigma_{\theta_x}}\phi(0), \tag{5.8}$$

and the corollary follows immediately from definition (5.3). □

Here is the interpretation of (5.7): first make the transformation $Y = g(X)$, which produces a location scale family $Y = \nu_\theta = \sigma_\theta\tilde{Z}$. Now apply (5.3) to this family. Since Y has density $\frac{1}{\sigma_\theta}\tilde{\phi}((y - \nu_\theta)/\sigma_\theta)$, (5.3) gives the transformation of Y, call it $h(y)$, with derivative at $y = g(x)$ equal to

$$h'(y) = 1/\sigma_{\theta_x}.$$

(Here we have used (2.13), and the fact that if $\text{Prob}_\theta\{X < x\} = .5$, i.e. if $\theta = \mu^{-1}(x)$, then $\text{Prob}_\theta\{Y < y\} = .5$, i.e., $\theta = \nu^{-1}(y)$.) According to (5.7), the transformation $\bar{g}(x)$ is the composition $hg(x)$.

In the case of an NSTF, $X = g^{-1}(\nu_\theta + \sigma_\theta Z)$, the transformed variable $Y = g(X)$ is perfectly normal, $Y \sim N(\theta, \sigma_\theta^2)$, but with nonconstant variance. Then (5.3) makes the further transformation $W = h(Y)$, where $h'(y) = 1/\sigma_y$, 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.

Notice that (5.7) can be rewritten as

$$\bar{g}(x_2) - \bar{g}(x_1) = \frac{g(x_2) - g(x_1)}{\sigma_{12}}, \quad x_1 < x_2,$$

where

$$\frac{1}{\sigma_{12}} \equiv \int_{x_1}^{x_2} \frac{1}{\sigma_{\theta_x}} g'(x)\, dy \Big/ \int_{x_1}^{x_2} g'(x)\, dx. \tag{5.9}$$

In an NSTF, $g(X) \sim N(\nu_\theta, \sigma_\theta^2)$, this has the following interpretation: $\bar{g}(x_2) - \bar{g}(x_1)$ is the number of standard deviations between $g(x_1)$ and $g(x_2)$, using the intermediate value σ_{12} as the unit of measurement. (Since $g'(x) > 0$, definition (5.9) necessarily makes σ_{12} intermediate between the extreme possible values of σ_{θ_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.

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

$$\mu_\theta : F_\theta(\mu_\theta) = .5.$$

Notice that the function μ_θ can be computed directly from the cdf's F_θ comprising $\mathscr{F}$, without any knowledge of the GSTF representation. This does not mean that it is easy to find a formula for μ_θ. In our examples the computations were done numerically.

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

$$\nu_\theta = g(\mu_\theta). \tag{6.1}$$

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 σ_θ is computed from (5.8),

$$\sigma_\theta = \frac{\phi(0) g'(\mu_\theta)}{f_\theta(\mu_\theta)}. \tag{6.2}$$

Notice that (3.1) can now be rewritten as

$$g(\mu_{\theta_0}) = 0, \qquad g'(\mu_{\theta_0}) = \frac{f_{\theta_0}(\mu_{\theta_0})}{\phi(0)}. \tag{6.3}$$

These two constraints complete the determination of the function g from formula (5.2). Comparing (6.3) with (5.3) shows that $g'(\mu_{\theta_0}) = \bar{g}'(\mu_{\theta_0})$.

For two values θ_1, θ_2, let $x_1 = \mu_{\theta_1}$ and $x_2 = \mu_{\theta_2}$. Then (5.2) and (6.2) together give

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

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

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

which is convenient for computation, especially in conjunction with (5.5).

NOTE. Formulas (6.1), (6.2) for ν_θ, σ_θ 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 θ_1, θ_2, simply by comparing $F_{\theta_1}(x)$ with $F_{\theta_2}(x)$ at different values of x, as in (1.6). Combined with (3.1), this gives σ_θ and ν_θ.

7. The continuous Poisson family. Figure 2 shows that the continuous Poisson family $\mathscr{F}$ is nearly an NSTF $X = g^{-1}(\nu_\theta + \sigma_\theta Z)$. Now we use the formulas of Sections 5 and 6 to calculate the functions g, ν_θ, and σ_θ, and also the variance stabilizing transformation $\bar{g}$, for $\mathscr{F}$. The results are shown in Table 2.

TABLE 2

Continuous Poisson family (2.9). The functions g, ν_θ, and σ_θ are calculated for the representation $X = g^{-1}(\nu_\theta + \sigma_\theta Z)$, under the constraints $\nu_{1.180} = 0$, $\sigma_{1.180} = 1$; also calculated is the variance stabilizing transformation $\bar{g}'$. Figures in parentheses relate to the traditional normalizing and variance stabilizing transformations. The constants $c_0 = .93$, $c_1 = 1.07$, $c_2 = 1.09$ are included to make the derivatives equal .93 at $x = 1$.

x	.125	.250	.5	1	2	4	8
$\theta = \mu^{-1}(x)$	.334	.450	.691	1.180	2.164	4.161	8.159
σ_θ	.85	.88	.93	1	1.09	1.19	1.29
ν_θ	−1.00	−.85	−.53	0	.85	2.14	4.15
$g'(x)$	1.45	1.29	1.12	.93	.75	.58	.46
$(c_0 x^{-1/3})$	(1.86)	(1.47)	(1.17)	(.93)	(.74)	(.59)	(.47)
$\bar{g}'(x)$	1.58	1.40	1.18	.93	.69	.49	.35
$(c_1(x+.33)^{-1/2})$	(1.59)	(1.41)	(1.18)	(.93)	(.70)	(.52)	(.37)
$(c_2(x+.375)^{-1/2})$	(1.54)	(1.38)	(1.17)	(.93)	(.71)	(.52)	(.38)

In Table 2, constraint (3.1) was applied with $\theta_0 = \mu^{-1}(1) = 1.180$; (6.3) then gives $g'(1) = .93$. The function $g'(x)$ was obtained from (5.5). (The stepwise algorithm described in the paragraph following (5.5) was applied with $\alpha = .45$.) The functions ν_θ and σ_θ were obtained from (6.1) and (6.2), respectively. Formula (5.3) gave $g'(x)$, also constrained to have $\bar{g}'(1) = .93$. Note that these calculations are valid assuming that $\mathscr{F}$ is an SSTF, not necessarily an NSTF.

Anscombe (1953) suggested $x^{2/3}$ as a normalizing transformation for the Poisson family, 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 $x \geq 1$, but, perhaps unsurprisingly, not for $x < 1$.

The variance stabilizing transformation $\sqrt{(x + .375)}$ derived by 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 $\sqrt{(x + b)}$ is obtained for $b = .33$. As a matter of fact $\sqrt{(x + .33)}$ stabilizes variances within the genuine Poisson family just as well as does $\sqrt{(x + .375)}$. Both transformations are superior to the naive transformation $\sqrt{x}$ in this regard. In this case, formula (5.3) has produced an excellent variance stabilizer.

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

$$\frac{\dfrac{d}{dx}\log \bar{g}'(x)}{\dfrac{d}{dx}\log g'(x)}. \tag{7.1}$$

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 $\mathscr{F}$, 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. One can show that any family $\mathscr{F}$ of continuous distributions which agrees with the Poisson family at the half-integer points, as at (2.10), must give almost the same results. Details appear in Section 7 of Efron (1981).

The story for the binomial family $Bi(n, \theta)$, n fixed, is similar to that for the Poisson. Figure 4 shows the diagnostic function $D(z, \theta)$ for the continuous Binomial family

$$F_\theta(x) = \int_\theta^1 t^{x-.5}(1-t)^{n-.5-x}\,dt\,\frac{\Gamma(n+1)}{\Gamma(x+.5)\Gamma(n-x+.5)}, \quad -½ < x < n + ½,$$

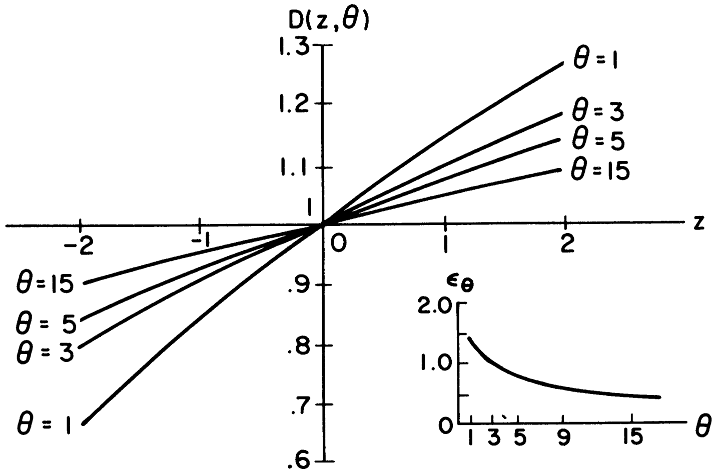

FIG. 4. *$D(z, \theta)$ for the continuous Binomial family, $n = 20$.*

$\theta \in (0, 1)$, $n = 20$. The cdf $F_\theta(x)$ equals that for the genuine binomial at the half-integer points $1/2, 3/2, \cdots, (n-1)/2$. Again we have nearly an NSTF family, with ε_θ large for θ 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 θ. That is, suppose that $\mathscr{F}$ is (nearly) an NSTF, 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 $\bar{g}(X)$, (5.3). 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.

For the rest of this section we consider the family (4.12), $X \sim N(\theta, (1 + \varepsilon\theta)^2)$ for $\theta > -1/\varepsilon$, ε a known constant. We have in mind values of ε in the range [0, 0.20]. Here $\mathscr{F}$ is already normal so $g(x) = x$. The variance stabilizing transformation (5.3) is

$$\bar{g}(x) = \frac{\log(1 + \varepsilon x)}{\varepsilon}.$$

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

$$W = \bar{\theta} + \bar{Z}, \tag{8.1}$$

where

$$\bar{\theta} = \bar{g}(\theta), \qquad \bar{Z} = \bar{g}(Z),$$

$Z \sim N(0, 1)$. Here we are ignoring the possibility $1 + \varepsilon Z < 0$, discussed in Section 4, an event with probability $\Phi(-1/\varepsilon)$, negligible for $\varepsilon \leq 0.2$.

Suppose that we want a central $1 - 2\alpha$ confidence interval for θ based on observing $X = x$. The obvious $1 - 2\alpha$ central interval for $\bar{\theta}$ based on observing $W = w$ in (8.1), is

$$\bar{\theta} \in [w - \bar{g}(z_{1-\alpha}), w - \bar{g}(z_\alpha)].$$

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

$$\theta \in \left[x + \hat{\sigma}\frac{z_\alpha}{1 - \varepsilon z_\alpha}, x + \hat{\sigma}\frac{z_{1-\alpha}}{1 - \varepsilon z_{1-\alpha}} \right], \tag{8.2}$$

where $\hat{\sigma} \equiv 1 + \varepsilon x$. For reasonable values of α, say $\alpha > .001$, interval (8.2) corresponds to inverting the locally most powerful tests in family (4.12), 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 (8.2) as the *true interval* for θ in what follows.

An easy approximate interval for θ, using normality but ignoring heteroscedasity, is

$$\theta \in [x + \hat{\sigma} z_\alpha,\ x + \hat{\sigma} z_{1-\alpha}]. \tag{8.3}$$

This is obtained by pretending that $\sigma_\theta = 1 + \varepsilon\theta$ is a constant in (4.12), and then estimating the constant by $\hat{\sigma} = 1 + \varepsilon x$.

Another approximate interval is obtained by ignoring the nonnormality of $\bar{Z}$ in (8.1). We suppose that

$$\bar{Z} \sim N(0, \beta). \tag{8.4}$$

Here we might take $\beta = 1$, since the transformation $\bar{g}$ is supposed to give unit variance, or we might use the actual variance of $\{\log(1 + \varepsilon Z)\}/\varepsilon$, some values of which are

(8.5)

$\varepsilon =$	0	.05	.1	.2
$\beta =$	1	1.0063	1.0261	1.1231

In either case, assumptions (8.1) and (8.4) lead to the interval

$$\theta \in \left[x + \hat{\sigma}\frac{\exp(\varepsilon\beta z_\alpha) - 1}{\varepsilon},\ x + \frac{\exp(\varepsilon\beta z_{1-\alpha}) - 1}{\varepsilon} \right]. \tag{8.6}$$

Table 3 compares the true interval (8.2) with (8.3) and with (8.6). The second approximation is seen to be better, more so if the correct value of β is used. In this highly simplified situation it is better to transform to homoscedasity and ignore nonnormality than vice-versa. Of course one could always do a complete analysis and recover the true interval (8.2), 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 $X_i \sim N(\theta_i, (1 + \varepsilon\theta_i)^2)$, $i = 1, 2, \cdots, n$. 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 (8.2). That is we will either fit the model

$$\boldsymbol{\theta} = \mathbf{M}\alpha \tag{8.7}$$

$\boldsymbol{\theta} = (\theta_1, \cdots, \theta_n)'$, $\mathbf{M}' = (m_1', m_2', \cdots, m_n')$, or the model

$$\bar{\boldsymbol{\theta}} = \mathbf{M}\alpha \tag{8.8}$$

$\bar{\boldsymbol{\theta}} = (\bar{\theta}_1, \cdots, \bar{\theta}_n)$, $\bar{\theta}_i = \bar{g}(\theta_i) = (1/\varepsilon)\log(1 + \varepsilon\theta_i)$. The fitting will be done by ordinary least squares (OLS) in either case, to $\mathbf{X} = \mathbf{x}$ in (8.7) or to $\mathbf{W} = \mathbf{w}$ in (8.8). The question is, which of these analyses will be asymptotically most efficient for estimating the unknown $k \times 1$ vector α, compared to maximum likelihood estimation? (Notice that there are two different maximum likelihood estimates, depending on whether we assume model (8.7) or (8.8). The efficiency comparisons are between OLS and the corresponding MLE for each of the two models.)

TABLE 3

Comparison of the true intervals (8.2) *with approximations* (8.3) *and* (8.6), *for* $\alpha = .05$, $x = 0$, $\hat{\sigma} = 1$. *Parenthetical numbers are one half the interval lengths.*

	$\varepsilon = .05$	$\varepsilon = .10$	$\varepsilon = .20$
True Interval (8.6)	[−1.520, 1.792] (1.656)	[−1.413, 1.969] (1.691)	[−1.238, 2.452] (1.845)
Approximate (8.7)	[−1.645, 1.645] (1.645)	[−1.645, 1.645] (1.645)	[−1.645, 1.645] (1.645)
Approximate (8.10), $\beta = 1$	[−1.579, 1.715] (1.647)	[−1.517, 1.788] (1.653)	[−1.402, 1.948] (1.675)
Approximate (8.10), true β	[−1.589, 1.726] (1.657)	[−1.553, 1.839] (1.696)	[−1.545, 2.235] (1.890)

TABLE 4
Comparison of the asymptotic efficiency of ordinary least squares on the variance stabilized scale, (8.9), *with the lower bound for efficiency on the normalized scale,* (8.10).

			ε			
			0	**.05**	**.1**	**.20**
		Eff_W	1	.9944	.9775	.9080
Lower Bound on Eff_X	**RATIO**	1	1	.9975	.9901	.9623
		1.5	.9798	.9774	.9701	.9428
		2	.9428	.9405	.9335	.9072
		3	.8660	.8638	.8575	.8333

Consider situation (8.8). The OLS estimate $\tilde{\alpha}$ and the MLE $\hat{\alpha}$ both have asymptotic covariance matrix of the form $c(\mathbf{M}'\mathbf{M})^{1/2}$. The ratio $c_{\hat{\alpha}}/c_{\tilde{\alpha}}$, which measures the asymptotic relative efficiency of $\tilde{\alpha}$ to $\hat{\alpha}$, say Eff_W, turns out to be

$$\mathrm{Eff}_W = 1/\sqrt{(1+2\varepsilon^2)\beta}, \tag{8.9}$$

β as given in (8.5); see Cox and Hinkley (1968).

Situation (8.7) is less neat, since in this case the OLS estimate $\tilde{\alpha}$ does not have a covariance matrix of form $c(\mathbf{M}'\mathbf{M})^{1/2}$. As a measure of efficiency comparable to (8.9), we take

$$\mathrm{Eff}_X = |\Sigma_{\hat{\alpha}}|^{1/(2k)}/|\Sigma_{\tilde{\alpha}}|^{1/(2k)},$$

where $\Sigma_{\hat{\alpha}}$ and $\Sigma_{\tilde{\alpha}}$ are the asymptotic covariance matrices. Using results of Bloomfield and Watson (1975), one can show that

$$\mathrm{Eff}_X \geq \frac{2\sqrt{\mathrm{RATIO}}}{1+\mathrm{RATIO}}, \qquad \mathrm{RATIO} \equiv \left[\frac{1+\varepsilon(\max_i \theta_i)}{1+\varepsilon(\min_i \theta_i)}\right]^2. \tag{8.10}$$

The lower bound (8.10) on Eff_X is achieved if the design matrix $\mathbf{M}$ has a special relationship to the covariance matrix of $\mathbf{X}$.

Table 4 compares Eff_W with the lower bound for Eff_X. If 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 ε, $|\varepsilon| \leq .1$, efficiency on the W scale cannot be much worse than 98%, which is quite safe indeed.

REFERENCES

ANSCOMBE, F. J. (1948). The transformation of Poisson, binomial, and negative binomial data. *Biometrika* **35** 246–254.

ANSCOMBE, F. J. (1953). Contribution to discussion on paper by H. Hotelling. *J. Roy. Statist. Soc. Ser. B* **15** 229–230.

BLOM, G. (1954). Transformations of the binomial, negative binomial, Poisson, and χ^2 distributions. *Biometrika* **41** 302–316.

BLOOMFIELD, P. and WATSON, G. (1975). The inefficiency of least squares. *Biometrika* **62** 129–132.

CRAMÉR, H. (1946) *Mathematical Statistics.* Princeton University.

COX, D. and HINKLEY, D. (1968). A note on the efficiency of least squares estimates. *J. Roy. Statist. Soc. Ser. B* **30** 284–289.

EFRON, B. (1981). Transformation theory: How normal is a family of distributions? Stanford University Technical Report No. 69.

FISHER, R. (1915). Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika* **10** 507–521.

HOTELLING, H. (1953). New light on the correlation coefficient and its transforms. (with discussion). *J. Roy. Statist. Soc. Ser. B* **15** 193–232.

JOHNSON, N. and KOTZ, S. (1970). *Continuous Univariate Distributions—2.* Houghton Mifflin, Boston.

TUKEY, J. (1957). On the comparative anatomy of transformations. *Ann. Math. Statist.* **38** 602–632.

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CA 94305