Open access • Journal Article • DOI:10.1093/BIOMET/62.1.101

On power transformations to symmetry — Source link

David Hinkley

Published on: 01 Apr 1975 - Biometrika (Oxford University Press)

Topics: Symmetry (physics) and Transformation (function)

ON POWER TRANSFORMATIONS TO SYMMETRY

By David Hinkley

School of Statistics

University of Minnesota

Technical Report No. 232

September 10, 1974

Summary.

Transformations to symmetry, or approximate symmetry, are considered. In particular, properties of simple estimates based on equitailed order statistics are derived. Examples include transformation of exponential and gamma random variables. Errors in previous work are discovered and partially corrected.

Some key words:

Symmetry; Transformation; Robustness; Maximum likelihood.

1. Introduction.

Box and Cox (1964) discussed estimation of data transformations which would yield variables satisfying a normal-error additive linear model. In particular, a family of power transformations was considered, which in its simple form consists of transformations $T_\lambda : y \to z_\lambda$ defined by

$$z_\lambda = \begin{cases} \dfrac{y^\lambda - 1}{\lambda} & \lambda \neq 0 , \\ \log y & \lambda = 0 . \end{cases} \tag{1.1}$$

Here y might be an observable quantity or a residual from a fitted model. A conventional assumption underlying the use of the transformation is that, for some λ, Z_λ has a normal distribution.

One method of estimating λ discussed by Box and Cox is that of maximum likelihood. This was further explored by Draper and Cox (1969), who derived expressions for the precision of the maximum likelihood estimate. Other aspects of normal-theory estimation and inference about λ in (1.1) have been investigated by Andrews (1971) and Atkinson (1973).

It is frequently assumed in connextion with (1.1) that y is positive; if y could be negative many values of λ would be clearly inadmissible. Note, however, that if y is positive then z_λ can have a normal distribution only if λ is zero or if λ^{-1} is an even integer. Nevertheless, one can often obtain a transformation for which Z_λ, although bounded below, is very nearly normal, or "close enough to normal for practical purposes."

There are three reservations that one might have about fitting (1.1) with a normal distribution assumption by maximum likelihood. First, the maximum likelihood method involves a great deal of calculation even in the normal case. Secondly, as Andrews (1971) has shown, the maximum likelihood method can be very sensitive to outliers; this reservation is actually unjustified in the sense that all reasonably efficient methods depend critically on the extreme observations. Thirdly, if we are aiming to use a linear model for the transformed data we may not want to make a normality assumption at any stage for fear of non-robustness. We may be planning to use now-popular robust methods of analysis (Huber, 1973), and the assumption of normality in connexion with (1.1) would seem contradictory.

In this paper we discuss simple and not-so-simple methods of estimating λ to give (approximate) symmetry for the distribution of Z_λ. The methods are based on symmetrizing order statistics about the median.

As were Draper and Cox (1969), we are not concerned here with the requirement of an additive linear model for transformed data. We do assume that the Y's have a common distribution with unknown location and scale.

In Section 2 we discuss a very simple order statistic estimate of λ and derive its large-sample properties. Corresponding results for the normal-theory maximum likelihood estimate are outlined in Section 3, which includes corrections to the results of Draper and Cox (1969). Section 4 then gives several illustrations of the results, for gamma, log normal and other distributions. Generalizations of the simple estimate of Section 2 are discussed in Section 5, with an example given in Section 6.

2. A Quick Estimate.

2.1. Definition of the Estimate.

Suppose that $Y_1,\ldots,Y_n$ are continuous non-negative independent and identically distributed random variables; the restriction to positive variables is necessary if the family (1.1) is to be sensible. If there exists a λ such that Z_λ in (1.1) has a symmetric distribution, then the p and $1-p$ quantiles will be symmetrically placed about the median. This symmetry of population quantiles for Z_λ suggests a simple method for estimating λ, namely that of symmetrizing the sample quantiles corresponding to tail probabilities p and $1-p$ for some p. As we suggested in the Introduction, Z_λ cannot have exact symmetry for most λ, but we assume that a value of λ exists which "nearly" gives symmetry. More will be said about this later.

Let $Y_1,\ldots,Y_n$ have the common distribution function $F(y)$, with quantiles ξ_s defined by

$$F(\xi_s) = s \quad 0 < s < 1 .$$

Then we week that transformation in the family (1.1) for which

$$\xi_{0.5}^{\lambda} - \xi_p^{\lambda} = \xi_{1-p}^{\lambda} - \xi_{0.5}^{\lambda} . \tag{2.1}$$

If we denote the ordered values of $Y_1,\ldots,Y_n$ by $X_1 \le X_2 \le \ldots \le X_n$, and define the median $\tilde{X}$ in the usual way, then the sample analog of (2.1) is

$$\tilde{X}^{\lambda} - X_r^{\lambda} = X_{n-r+1}^{\lambda} - \tilde{X}^{\lambda} \quad r = [np], \tag{2.2}$$

which is an estimating equation for λ. There are only two solutions to (2.2), one of them being $\lambda = 0$. However, by comparison with (1.1) we exclude $\lambda = 0$ unless

$$\frac{\tilde{X}}{X_r} = \frac{X_{n-r+1}}{\tilde{X}} \tag{2.3}$$

which is the condition for sample quantiles of $\log Y$ to be symmetric about the median.

For computation purposes it is easier to rewrite (2.2) in the form

$$\left(\frac{X_r}{\tilde{X}}\right)^{\lambda} + \left(\frac{X_{n-r+1}}{\tilde{X}}\right)^{\lambda} = 2 \ . \tag{2.4}$$

The existence of one non-zero solution to (2.4) is easily proved directly, or as a special case of the lemma in Section 5. The non-zero solution T of (2.4) is positive if and only if

$$X_r \cdot X_{n-r+1} > \tilde{X}^2$$

and is otherwise negative if (2.3) is not satisfied; this is obviously sensible on physical grounds. Moreover it is easy to verify that

$$|T| > |\log \log(X_{n-r+1}/\tilde{X}) - \log \log(\tilde{X}/X_r)| \div |\log(X_r/X_{n-r+1})$$

which may be useful in solving (2.4).

The estimator defined by (2.3) and (2.4) is somewhat naive. One would expect that in order to obtain a reasonably efficient estimator one would have to combine the equations (2.2) corresponding to several p values in some sensible way. This we do in Section 5. However the simplicity of (2.2) is appealing, and there is some flexibility in our ability to choose p. Also the reasonableness of the basic idea and some generally useful properties of the estimator are most easily discussed in the simple case.

2.2. Properties of the Quick Estimate .

We have already seen that T, the non-zero estimator satisfying (2.4), is unique. We now show that as $n \to \infty$ the estimator T defined by (2.4) has a limiting normal distribution. To do this we use the joint asymptotic normality of the order statistics $X_{r_1},\ldots,X_{r_m}$ for $r_j = [np_j]$, $0 < p_1 <\ldots< p_m < 1$. Specifically, if the original distribution function $F(y)$ has density $f(y)$ and quantiles $\xi_s = F^{-1}(s)$, then the vector $(X_{r_1},\ldots,X_{r_m})$ has a limiting multivariate normal distribution with mean $(\xi_{p_1},\ldots,\xi_{p_m})$, and covariance matrix determined by

$$n\,\mathrm{cov}(X_{r_i},X_{r_j}) = \frac{p_i(1-p_j)}{f(\xi_{p_i})f(\xi_{p_j})}\ , \quad i \le j\ . \tag{2.5}$$

The first property of the estimator T that we need is consistency, which strictly means that

$$T = \lambda + o_p(1)$$

where λ is the solution of (2.1); if there is a transformation in the class (1.1) giving exact symmetry, then the solution to (2.1) gives it, whatever p. Actually consistency is easy to verify from continuity of the left-hand side of (2.4) and consistency of X_r, X_{n-r+1} and $\tilde{X}$ for the respective quantiles. Note that $\tilde{X}$ is asymptotically equivalent to $X_{[\frac{1}{2}n]}$, which fact we shall use.

Now let us suppose $\lambda \neq 0$, and write

$$X_r = \xi_p(1 + n^{-\frac{1}{2}} W_p), \quad X_{n-r+1} = \xi_{1-p}(1 + n^{-\frac{1}{2}} W_q), \quad r = [np], \; p+q = 1,$$

and

$$\tilde{X} = \xi_{0.5}(1 + n^{-\frac{1}{2}} W_{0.5}) . \tag{2.6}$$

Then the estimating equation (2.4) can be written

$$[\alpha_p\{1+n^{-\frac{1}{2}}(W_p-W_{0.5}) + o_p(n^{-\frac{1}{2}})\}]^T + [\alpha_q\{1+n^{-\frac{1}{2}}(W_q-W_{0.5}) + o_p(n^{-\frac{1}{2}})\}]^T = 2, \tag{2.7}$$

where $\alpha_s = \xi_s/\xi_{0.5}$ and the α's satisfy

$$\alpha_p^\lambda + \alpha_q^\lambda = 2 \tag{2.8}$$

by definition. Since T is consistent, expansion of (2.7) about $T = \lambda$ gives, using (2.8),

$$(T-\lambda)(\alpha_p^\lambda \log \alpha_p + \alpha_q^\lambda \log \alpha_q) + \lambda n^{-\frac{1}{2}}\{\alpha_p^\lambda(W_p-W_{0.5}) + \alpha_q^\lambda(W_q-W_{0.5})\} + o_p(T-\lambda)$$
$$+ o_p(n^{-\frac{1}{2}}) = 0 .$$

That is, to first order,

$$\sqrt{n}(T-\lambda)/\lambda = \frac{2W_{0.5} - \alpha_p^\lambda W_p - \alpha_q^\lambda W_q}{\alpha_p^\lambda \log \alpha_p + \alpha_q^\lambda \log \alpha_q} \tag{2.9}$$

We then use the limiting joint normality of the W's, whose covariance matrix is determined by (2.5) and the transformation (2.6), to obtain the limiting normal distribution of T. If we define $h_s^{-1} = \xi_s f(\xi_s)$, the variance of the limiting normal distribution of $\sqrt{n}(T-\lambda)$ is found to be

$$V_T(\lambda,p) = \frac{\lambda^4\{h_{\frac{1}{2}}^2 + pq(\alpha_p^{2\lambda} h_p^2 + \alpha_q^{2\lambda}\, h_q^2) - qp(\alpha_p^\lambda\, h_p + \alpha_q^\lambda\, h_q)h_{\frac{1}{2}} + 2p^2\alpha_p^\lambda\, \alpha_q^\lambda\, h_p\, h_g\}}{(\alpha_p^\lambda \log \alpha_p^\lambda + \alpha_q^\lambda \log \alpha_q^\lambda)^2} \qquad (2.10)$$

An alternative expression, in terms of the p.d.f. $g(z)$ for the transformed variable Z_λ, is

$$V_T(\lambda,p) = \frac{\lambda^4\{g_{\frac{1}{2}}^{-2} + pq(g_p^{-2} + g_q^{-2}) - 2p(g_p^{-1}\, g_{\frac{1}{2}}^{-1} + g_q^{-1}\, g_{\frac{1}{2}}^{-1}) + 2p^2 g_p^{-1}\, g_q^{-1}\}}{\{(1+\lambda K_p)\log(1+\lambda K_p) + (1+\lambda K_q)\log(1+\lambda K_q) - 2(1+\lambda K_{\frac{1}{2}})\log(1+\lambda K_{\frac{1}{2}})\}^2}, \qquad (2.11)$$

where K_s is the quantile defined by $G(K_s) = s$ and $g_s = g(K_s)$.

Notice that the properties of T are invariant under scale change of Y, as is immediately obvious from the estimating equation (2.4).

The above results hold also for $\lambda = 0$, when $Z = \log Y$. Slightly more generally, (2.11) for small λ may be written as

$$V_T(\lambda,p) = \frac{g_{\frac{1}{2}}^{-2} + pq(g_p^{-2} + g_q^{-2}) - 2pg_{\frac{1}{2}}^{-1}(g_p^{-1} + g_q^{-1}) + 2p^2 g_p^{-1}\, g_q^{-1}}{\{\frac{1}{2}(K_p^2 + K_q^2 - 2K_{\frac{1}{2}}^2) - \frac{1}{6}\lambda(K_p^3 + K_q^3 - 2K_{\frac{1}{2}}^3)\}^2}. \qquad (2.12)$$

This contradicts the type of result obtained by Draper and Cox, but their results are wrong as we show in the next section.

Several examples illustrating the results of this section are given in Section 4.

3. Normal-theory Maximum Likelihood.

As we pointed out in the introduction, previous work on power transformations has assumed the transformed variable Z_λ to be normally distributed; in the simplest case the variables are taken to be homogeneous $N(\mu,\nu)$. Draper and Cox derived large-sample properties of the estimator $\hat{\lambda}_N$ obtained by maximizing the $N(\mu,\nu)$ likelihood. These properties would provide useful standards by which to judge the simple estimate T described in Section 2; however some of the Draper and Cox results are incorrect and others are incomplete. We therefore briefly outline the basic properties of the normal-theory maximum likelihood estimate $\hat{\lambda}_N$ here.

The $N(\mu,\nu)$ likelihood e^L for $Z_{\lambda,1},\ldots,Z_{\lambda,n}$ leads directly to the efficient score vector $U.$ given by

$$U_{\lambda.} = \frac{\partial L}{\partial \lambda} = \Sigma \log Y_j - \nu^{-1}\lambda^{-2}\, \Sigma\, (Z_j-\mu)\{(1+\lambda Z_j)\log(1+\lambda Z_j) - \lambda Z_j\}$$

$$U_{\mu.} = \frac{\partial L}{\partial \mu} = \nu^{-1}\, \Sigma\, (Z_j-\mu)$$

$$U_{\nu.} = \frac{\partial L}{\partial \nu} = (2\nu^2)^{-1}\, \Sigma\, (Z_j-\mu)^2 - (2\nu)^{-1}. \qquad (3.1)$$

An obvious feature of the component likelihood equation $U_{\lambda.} = 0$ is its invariance under scale transformation of the original variable Y.

Provided that the density $f(y)$ of Y is regular and a unique solution of $E(U.) = 0$ exists, as is the case for standard continuous distributions on $[0, \infty)$, the normal-theory maximum likelihood estimate converges stochastically to the solution of $E(U.) = 0$ and has a limiting normal distribution.

Let $\theta = (\lambda, \mu, \nu)$ and denote the normal-theory m.l.e. by $\hat{\theta}_N$ with limit θ_N. A standard expansion of the likelihood equation gives

$$\sqrt{n}(\hat{\theta}_N - \theta_N) = \{- \frac{1}{n} \frac{\partial^2 L}{\partial \theta^2} \Big|_{\theta=\theta_N} \}^{-1} \frac{1}{\sqrt{n}} U.(\theta_N) + o_p(1); \tag{3.2}$$

see, for example, Cox and Hinkley (1974, Chapter 9). Then $\sqrt{n}(\hat{\theta}_N - \theta_N)$ has a limiting normal distribution with covariance matrix

$$\underset{\sim}{\Sigma} = \underset{\sim}{J}^{-1} \underset{\sim}{I} \underset{\sim}{J}^{-1} , \tag{3.3}$$

where

$$n\underset{\sim}{J} = E_f(- \frac{\partial^2 L}{\partial \theta^2} \Big|_{\theta=\theta_N})$$

and

$$n\underset{\sim}{I} = E_f\{U.(\theta_N)U.'(\theta_N)\} ; \tag{3.4}$$

here E_f denotes expectation with respect to the density $f(y)$ of Y. Note that $\underset{\sim}{\Sigma} = \underset{\sim}{I}^{-1}$ only if f is the normal density because $I = J$ only if L is the log likelihood according to the density f. The general form (3.3) is required when examining properties of $\hat{\lambda}_N$ under non-normal distributions, as we do in Section 4.

Draper and Cox incorrectly obtain the variance of $\hat{\lambda}_N$ from I^{-1}. Their method of expanding $U.$ as a power series in λ does lead to approximations for $\underset{\sim}{I}$ and $\underset{\sim}{J}$ up to any order in λ, but the results for $\underset{\sim}{\Sigma}$ are very complicated (involving the first six moments of Z_λ) and of limited usefulness. In particular cases one can evaluate $\underset{\sim}{\Sigma}$. Some general results for the case $\lambda = 0$ are given in Section 4.

4. EXAMPLES

4.1. Exponential and gamma cases.

To illustrate the discussion up to this point we first examine in some detail the example chosen by Draper and Cox, where the original variables $Y_1,\ldots,Y_n$ are exponentially distributed with density

$$f(y) = \rho \exp(-\rho y).$$

In this particular case (2.2) becomes

$$(-\log p)^\lambda + (-\log q)^\lambda = 2(\log 2)^\lambda, \ p + q = 1. \qquad (4.1)$$

The quantiles of Y^λ are

$$\eta_s(\rho,p) = \rho^{-\lambda} \{-\log(1-s)\}^\lambda , \qquad (4.2)$$

and the quantiles $K_s(\rho,p)$ of Z_λ are given by $K = (\eta-1)/\lambda$. A crude outlier-free measure of asymmetry for Z_λ is the "tilt factor"

$$\tau(s,p) = \frac{\eta_{1-s}(\rho,p) - \eta_{0.5}(\rho,p)}{\eta_{0.5}(\rho,p) - \eta_s(\rho,p)}, \quad 0 < s < \tfrac{1}{2} . \qquad (4.3)$$

Note that the non-zero solution λ_p of (4.1) and $\tau(s,p)$ are both independent of the scale parameter.

Table 1 gives some values of λ_p and $\tau(s,p)$ for $p, s \geq 0.01$. The entries show that λ_p is very nearly constant for $p > 0.10$; and, related to this stability, there is a high degree of symmetry as far as the upper and lower 5% points of the transformed distributions. Much the same conclusions were reached by Draper and Cox, who noted that small changes in λ have little visible effect on the symmetry. The Weibull distribution of Z_λ is quite close to normal except in the extreme tails.

Table 1. Transformations and tilt factors in the exponential case.

Quantile p		.005	0.01	0.05	0.10	0.20	0.30	0.40
Transformation power λ_p		.272	0.28	0.291	0.297	0.303	0.305	0.307
	s=0.2	0.970	0.978	0.989	.995	1.000	1.002	1.004
Tilt factor	s=0.1	0.963	0.975	0.991	1.000	1.009	1.018	1.015
$\tau(s,p)$	s=0.05	0.964	0.979	1.000	1.011	1.023	1.027	1.031
	s=0.02	0.973	0.992	1.019	1.034	1.047	1.054	1.059
	s=0.01	0.985	1.007	1.038	1.055	1.072	1.078	1.084

The limiting normal distribution of T is scale invariant, as we noted in Section 2, and hence independent of ρ. The variance V_T is given in Table 2 for the same transformations described in Table 1; rows below that for the exponential case are defined later.

Table 2. Large-sample variance V_T of the quantile transformation estimate for exponential and gamma distributions.

r \ p	.005	0.01	0.05	0.10	0.20
1, exponential	0.589	0.582	1.012	1.894	6.271
2	1.704	1.670	2.968	6.148	19.916
3	2.841	2.718	4.507	8.982	36.069
4	3.977	3.748	5.936	11.442	43.420

It is interesting to see that rather extreme order statistics give the best precision, $p = .01$ being close to optimal. This is a pity, in a sense, because rather large samples would be required for anyone to have faith in the results! Also the method is consequently sensitive to outliers.

The corresponding results for the normal-theory m.l.e. $\hat{\lambda}_N$ are easily derived using the efficient score formulae in (3.1) together with the identity

$$\int_0^\infty (\log y)^r \, y^s \, e^{-y} \, dy = \frac{d^r}{ds^r} \Gamma(1+s) \qquad s \geq 0,$$

which is related to the polygamma functions. The maximum likelihood estimate $\hat{\lambda}_N$ converges to 0.265 (cf. Draper and Cox's approximation 0.268), and

$$\hat{\mu}_N \to \rho^{-\lambda_N} \Gamma(1+\lambda_N), \; \hat{\nu}_N + \hat{\mu}_N^2 \to \rho^{-2\lambda_n} \Gamma(1+2\lambda_N) \; .$$

The variance V_N of the limiting normal distribution of $\sqrt{n}(\hat{\lambda}_N - \lambda_N)$ is 0.314. Note from Table 1 that $\lambda = 0.265$ gives a relatively poor degree of symmetry.

The above calculations for the exponential case are easily extended to the general gamma density

$$f(y) = y^{r-1} \, e^{-y} \div \Gamma(r) \; ,$$

and we have added such calculations in Tables 2 and 3 for $r = 2,3$ and 4. The correct transformation power λ_p for T is quite stable at about 0.32 for these cases, i.e., close to the conventional cube root transformation.

As r increases, the transformed variable Z_λ is closer to symmetry (and normality).

Table 3. Large-sample limit λ_N and variance V_N of the normal-theory MLE of λ for exponential and gamma distributions.

r	1, exponential	2	3	4
λ_N	0.2654	0.301	0.312	0.318
V_N	0.314	0.914	1.567	2.229

4.2. Examples with $\lambda = 0$.

For the special case $\lambda = 0$ equation (2.12) gives a simple expression for V_T, the large-sample variance of $\sqrt{n}(T-\lambda_p)$. A corresponding result for the normal-theory maximum likelihood estimate is quite easily derived from (3.3). Lengthy algebra gives

$$V_N = \frac{36(\nu^2\mu_{(6)} - 6\nu^3\mu_{(4)} - 2\nu\mu_{(3)}\mu_{(5)} + \mu_{(3)}^2\mu_{(4)} + 7\nu^2\mu_{(3)}^2 + 9\nu^5)}{(7\nu\mu_{(4)} - 6\mu_{(3)}^2 - 3\nu^3)^2}, \quad (4.4)$$

where $\mu_{(r)}$ is the r^{th} central moment of $Z_0 = \log Y$. We now look at two specific examples.

When $\log Y$ has the $N(\mu,\nu)$ density, (2.12) and (4.4) simplify to

$$V_T = x_p^{-4}\nu^{-1}(\phi_{0.5}^{-2} + 2p\phi_p^{-2} - 4p\phi_p^{-1}\phi_{0.5}^{-1}), \quad (4.5)$$

where $\Phi(x_s) = s$ and $\phi_s = \phi(x_s)$, and

$$V_N = \frac{2}{3}\nu^{-1}.$$

Some numerical values of V_T are given in Table 4. The smallest value of V_T occurs at $p = 0.01$, at which point $V_N/V_T \doteq 2/\pi$, rather interestingly.

Table 4. Large-sample Variances V_T for quantile transformation estimate in log normal and log double exponential cases.

p	0.005	0.01	0.02	0.05	0.10	Normal-theory maximum likelihood
Normal: νV_T	1.15	1.04	1.08	1.48	2.62	0.667
Double exponential: $\rho^{-2}V_T$	0.881	0.837	0.894	1.28	2.39	1.491

Note: The variances of Y are respectively ν and $2\rho^{-2}$.

The effect of unknown λ on estimation of μ and ν is seen from the complete covariance matrix

$$\Sigma_N = n\ \mathrm{var}(\hat{\theta}_N) = \begin{bmatrix} \frac{2}{3}\nu^{-1} & \frac{\nu+\mu^2}{3\nu} & \frac{4}{3}\mu \\ \cdot & \nu + \frac{(\nu+\mu^2)^2}{6\nu} & \underline{2\mu(\nu+\mu^2)} \\ \cdot & \cdot & 2\nu^2 + \frac{8}{3}\mu^2\nu \end{bmatrix} .$$

The potentially heavy increase in $\mathrm{var}(\hat{\mu})$ due to not knowing λ is clearly worth investigating in more generality.

If $\log Y$ has a distribution close to the normal, so that the standardized moments $\gamma_1 = \mu_{(3)}/\nu^{3/2}$, $\gamma_2 = \mu_{(4)}\nu^{-2}-3$, etc. are of successively lower order in some notional parameter, we can approximate V_N from (4.4) by

$$V_N = \frac{2}{3}\nu^{-1}\left(1 - \frac{16}{9}\gamma_2 + \frac{11}{6}\gamma_1^2\right) .$$

In a sense this corresponds to (9) of Draper and Cox, their factor θ^2 being incorrect.

A corresponding approximation for V_T is easily constructed from (2.12) using a Fisher-Cornish expansion for K_s and an Edgeworth expansion for $g(z)$. The result is somewhat complicated and will not be given here.

A distribution characterising much longer tails than the normal is the double exponential, with density

$$g(z) = \frac{1}{2}\rho\ \exp(-\rho|z|) \qquad -\infty < z < \infty .$$

If $\log Y$ has this distribution, it is easy to show that (2.12) becomes

$$V_T = \rho^2(\log 2p)^{-4}(2p^{-1} - 4) \qquad 0 < p < \frac{1}{2} .$$

with values as in Table 4. The corresponding value of V_N calculated from (4.4) is $1.491\rho^2$ so that T is superior to $\hat{\lambda}_N$ in large samples for $p \le .06$. In terms of the variance ν of Z, the smallest value of V_T here is $1.674\nu^{-1}$, compared to $1.044\nu^{-1}$ in the log-normal case.

5. GENERALIZATION OF THE QUICK ESTIMATE.

5.1. The generalization.

There are several ways in which one could generalize the estimator T defined by (2.2). First, we could solve (2.2) for several values of p and average the resulting estimates of θ. Secondly, we could, as it were, average the equation (2.2) for several p values and then solve for the estimator. Other possible methods exist, but this latter method is the one we examine here.

We propose, then, to use the equation (2.2) for several values of p, say $p_1 < \ldots < p_m < 1/2$, and in fact to form the combined equation

$$\sum_{j=1}^{m} c_j (X_{r_j}^T + X_{n-r_j+1}^T) = 2 \sum_{j=1}^{m} c_j \tilde{X}^T \tag{5.1}$$

where $r_j = [np_j]$; the solution $T = 0$ is chosen only if

$$\sum c_j \log(X_{r_j} X_{n-r_j+1}) = 2 \sum c_j \log(\tilde{X}) , \tag{5.2}$$

corresponding to (2.3). The coefficients $c_1, \ldots, c_m$ in (5.1-2) are arbitrary weights to be chosen. A more convenient form of (5.1) is

$$\sum c_j \left\{ \left(\frac{X_{r_j}}{\tilde{X}}\right)^T + \left(\frac{X_{n-r_j+1}}{\tilde{X}}\right)^T \right\} = 2 \sum c_j . \tag{5.3}$$

In practice it would be sensible to choose all c_j's positive, particularly if a monotone transformation of Y is symmetrically distributed, since otherwise asymmetry of quantile pairs tends to cancel out in the summation.

The existence of a unique non-zero solution to (5.3) for positive c_j is easily proved by the following lemma.

Lemma. For arbitrary positive constants $a_1,\dots,a_m$, $b_1,\dots,b_m$ and $c_1,\dots,c_m$, the equation

$$\sum c_j(a_j^t + b_j^t) = 2\sum c_j \tag{5.4}$$

has a single non-zero real solution unless

$$\sum c_j \log(a_j b_j) = 0$$

in which case $t = 0$ is the only solution.

Proof is obvious by defining a random variable U with values $\log a_j$ and $\log b_j$ $(j = 1,\dots,m)$, and probabilities $c_j/(2\sum c_i)$ at $U = \log a_j$ and $\log b_j$. Then (5.4) is the equation

$$E(e^{tU}) = 1 \, , \tag{5.5}$$

which has a unique non-zero solution unless $E(U) = 0$. (It is interesting to note that the strong ordering $a_1 < \dots < a_m < 1 < b_m < \dots < b_1$ is not used here, suggesting that a stronger result holds for T.)

A useful and obvious corollary of the representation (5.5) is that T is negative (positive) if the left side of (5.2) is greater than (less than) the right side.

Although the general equation (5.2) is interesting theoretically for any value of m, in practice one might well restrict attention to $m = 2$ or 3 and use equal weights c_j. Potentially the use of $m > 1$ could accomplish two things: (i) increased precision of the transformation estimate, (ii) an averaging out of the asymmetry in Z_T when no Z_λ has a symmetric distribution.

5.2. Large-sample properties.

The groundwork for establishing large-sample properties of T as been laid in Section 2.2. Here we outline the main steps and results.

By continuity of (5.2) and consistency of the order statistics, T is consistent for that value λ_p of λ satisfying

$$\sum c_j(\xi_{p_j}^{\lambda} + \xi_{q_j}^{\lambda}) = 2\sum c_j\, \xi_{0.5}^{\lambda} ,$$

which would be common to all p if Z_λ is symmetrically distributed. By the same expansion route used in Section 2.2 we find that for all λ

$$\sqrt{n}(T-\lambda) \doteq \lambda = \frac{2W_{0.5} - \sum c(\alpha_p^{\lambda} W_p + \alpha_q^{\lambda} W_q)}{\sum c(\alpha_p^{\lambda} \log \alpha_p + \alpha_q^{\lambda} \log \alpha_q)} + o_p(1) . \tag{5.6}$$

(Here and below the suffix j on c_j, p_j and q_j has been dropped for typographical convenience.) The resulting limiting normal distribution for T is again obtained from the limiting joint normal distribution of order statistics, and using (2.5) the variance $V_T(\lambda,p)$ is found to be equal to

$$\begin{aligned} V_T(\lambda,p) = \lambda^4[h_{0.5}^2 - 2h_{0.5}\sum cp(\alpha_p^{\lambda} h_p + \alpha_q^{\lambda} h_q) + \sum c^2\{pq(\alpha_p^{2\lambda} h_p^2 + \alpha_q^{2\lambda} h_q^2) \\ + 2p^2 \alpha_p^{\lambda} \alpha_q^{\lambda} h_p h_q\} + 2\sum_{p<p'} cc' \{pq'(\alpha_p^{\lambda} \alpha_{p'}^{\lambda} h_p h_{p'} + \alpha_q^{\lambda} \alpha_{q'}^{\lambda} h_q h_{q'}) \\ + pp'(\alpha_p^{\lambda} \alpha_{q'}^{\lambda} h_p h_{q'} + \alpha_{p'}^{\lambda} \alpha_q^{\lambda} h_{p'} h_q)\}] \\ \div \{\sum c(\alpha_p^{\lambda} \log \alpha_p + \alpha_q^{\lambda} \log \alpha_q)\}^2 . \end{aligned} \tag{5.7}$$

The notation throughout is that of Section 2.2.

A corresponding expression for V_T in terms of the p.d.f. $g(z)$ can be obtained from (5.7) in the same way that (2.11) was derived from (2.10). This simply amounts to substituting g_s^{-1} for $\alpha_s^{\lambda} h_s$ in the numerator and $(1+\lambda K_s)\log(1+\lambda K_s) - (1+\lambda K_{0.5})\log(1+\lambda K_{0.5})$ for $\alpha_s^{\lambda} \log \alpha_s$ in the denominator of (5.7), where we recall that $G(K_s) = s$ and $g_s = g(K_s)$.

The result (5.7) as we have given it is for finite m, and would apply when m is small relative to n. If all the order statistics X_j are used, so that $m = [\frac{1}{2}n]$ in (5.1), a corresponding asymptotic result can be obtained for a smooth weight function $c(x)$ defined by

$$c_j = c(\frac{j}{n+1}) .$$

In terms of the p.d.f. $g(z)$ the result is

$$\lambda^{-4} V_T(\lambda, c) = \frac{\{\psi(\frac{1}{2})\}^2 - 2\psi(\frac{1}{2})A_1(c) + A_2(c) + A_3(c)}{\{B(c)\}^2} ,$$

where $\psi(x) = 1/g\{G^{-1}(x)\}$ and

$$A_1(c) = \int_0^1 c(x)x\{\psi(x) + \psi(1-x)\}dx ,$$

$$A_2(c) = \int_0^1 x(1-x)c^2(x)\{\psi^2(x) + \psi^2(1-x)\}dx ,$$

$$A_3(c) = \int\int_{x<x'} c(x)c(x')[x(1-x')\{\psi(x)\psi(x') + \psi(1-x)\psi(1-x')\}$$

$$+ xx'\{\psi(x)\psi(1-x') + \psi(1-x)\psi(x')\}]dxdx'$$

and

$$B(c) = \int_0^1 c(x)[\{1+\lambda G^{-1}(x)\}\log\{1+\lambda G^{-1}(x)\}+\{1+\lambda G^{-1}(1-x)\}$$

$$\log\{1+\lambda G^{-1}(1-x)\}]dx-2\{1+\lambda G^{-1}(\tfrac{1}{2})\}\log\{1+\lambda G^{-1}(\tfrac{1}{2})\} .$$

A discussion of conditions required for this result will not be given here; a recent reference is the paper by Stigler (1974).

6. AN EXAMPLE.

After introducing the generalization of T in Section 5, we need to assess what is gained in precision at the expense of complication. From calculations we have done it would seem that little is to be gained using the generalization. Here we give only one example, the case where $\log Y$ is normally distributed.

When $\lambda = 0$ and $Z = \log Y$ is $N(\mu, \nu)$, we saw in Section 4.2 that T has minimum large-sample variance at $p = 0.01$, where $V_T = 1.04\nu^{-1}$. Using a simplified form of (5.7) corresponding to (2.12), we obtain the results given in Table 4. The right hand column of the table gives values of νV_T, and the other entries indicate values of c_j; the c_j's sum to one in each case.

TABLE 4.

Large-sample variance V_T of the generalized version of T when $\log Y$ is $N(\mu, \nu)$.

p	0.01	0.02	0.05	0.10	νV_T
	1	0	0	0	1.04
	½	½	0	0	0.92
	0	1	0	0	1.08
	½	0	½	0	0.91
	0	0	1	0	1.48
c	0	½	½	0	1.04
	0	0	2/3	1/3	1.54
	0	0	½	½	1.64
	0	0	0	1	2.62
	1/3	1/3	1/3	0	0.88
	0	1/3	1/3	1/3	1.12

Some general features are apparent from this small set of results. Most striking is the fact that if all values of p exceed .05, then $m = 1$ (one pair of order statistics) cannot be markedly improved on by $m = 2$. Use of $m = 3$ with one value of p equal to .01 can give up to 15% improvement in precision, which is a little better than using $m = 2$. With $p = .05, .10, .15,$ and $.20$ and each c_j equal to $1/4$, $\nu V_T = 2.23$. We conclude that it is not possible to escape the extreme tails ($p \leq .02$) and keep precision, unless perhaps m is considerably larger than 3.

7. FURTHER DISCUSSION.

Use of power transformations such as (1.1) occurs most frequently with more complicated linear models than the single mean case discussed in this paper. The ability to generalize the estimator T defined by (5.1) depends to some extent on whether or not the linear model design includes replication.

Suppose that Y_{ij}, $j = 1,\ldots,r_i$, are replicates of the i^{th} cell of a linear model, meaning that for some λ

$$Z_{\lambda,ij} = \mu_i + e_{ij} \ . \tag{7.1}$$

We can generalize (2.2) and (2.3), or (5.1) and (5.2), as follows. Let $\tilde{Y}_i$ be the median of variables in the i^{th} cell, and define

$$A_{ij} = Y_{ij}/\tilde{Y}_i \ , \qquad j = 1,\ldots,r_i, \ \ i = 1,\ldots,I. \tag{7.2}$$

Then the ordered values of A_{ij} replace the ratios $X_i/\tilde{X}$ in (5.1) and (5.2). The estimating equation so defined is not a trivial generalization, although the consistency of T for fixed I and large $n = \Sigma r_i$ is still assured. The problem is that the standardization in (7.2) is non-homogeneous, the more so if the variability of μ_i is large relative to that of the e_{ij} in (7.1). Assuming that the e_{ij} are homogeneous errors, it is clear that if

$$\mathrm{var}(Y_{ij}) \propto \{E(Y_{ij})\}^b \ ,$$

then cells with larger means will dominate the estimating equation, and hence T, if $b > 2$. For example, if $\lambda = 1$ in (7.1) then $b = 0$ and cells with small means dominate T; if $\lambda = 0$ then $b = 2$ and no cell dominates T.

While we have not examined this problem in any detail, this does seem to be a suitable situation for use of the generalization (5.1) with $m = [\frac{1}{2}n]$ and c_j = constant. This has the disadvantage of requiring a large amount of computation.

An example that fits into this discussion is the first numerical example in Box and Cox (1964), which is a fourfold replicate of a 3×4 design. The normal-theory likelihood suggests that $\lambda = -1$, although one would not discount values $-1 < \lambda < 0$. The three outmost pairs of ordered A_{ij}'s each yield the estimate $T = 0$ by the method of Section 2. Fitting the additive two-way linear model by least squares with $\lambda = -1$ and $\lambda = 0$ gives negligible interactions. Normal plots of residuals reveal that $\lambda = -1$ gives a better fit to normality, although the closeness to symmetry is about the same for both $\lambda = -1$ and $\lambda = 0$; in each case there are two or three moderately large outliers (not the same data points). There is some evidence that extreme A_{ij}'s are associated with large cell means, which suggests that λ is somewhat negative. Strangely, use of less extreme A_{ij}'s indicates λ to be around 0.5 although there is no consistent value for any particular pair.

This discussion is intended to suggest that there are difficulties with the order-statistic method, particularly in connexion with complex models. When one is able to use the simple estimating equation (2.2), either in the original form or with the A_{ij} defined in (7.2), the estimate T should be reasonably constant over the outermost pairs of order statistics in order to be convincing. It would be helpful to understand more clearly the problem of heterogeneity in the A_{ij}'s, particularly through experience with applications.

One must conclude, however, that the need to use fairly extreme order statistics in order to achieve precise estimates of λ makes the quick method of Section 2 unappealing with moderate amounts of data containing genuine outliers. It would seem that data transformation in the presence of outliers is a risky business.

References

Andrews, D. F. (1971). A note on the selection of data transformations. Biometrika 58, 249-54.

Atkinson, A. C. (1973). Testing transformations to normality. J. R. Statist. Soc. B 35, 473-9.

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

Cox, D. R. and Hinkley, D. V. (1974). Theoretical Statistics. New York: Wiley; London: Chapman-Hall.

Draper, N. R. and Cox, D. R. (1969). On distributions and their transformation to normality. J. R. Statist. Soc. B 31, 472-6.

Huber, P. J. (1973). Robust regression: Asymptotics, conjectures and Monte Carlo. Ann. Statist. 1, 799-821.

Stigler, S. M. (1974). Linear functions of order statistics with smooth weight functions. Ann. Statist. 2,676-93.