Analysis of Transformed Data*

by

D.V. Hinkley & G. Runger
University of Minnesota
School of Statistics
Technical Report No. 341
July 1980

*Research supported by NSF Grant MCS-7904558

ANALYSIS OF TRANSFORMED DATA.

BY

D. V. HINKLEY* AND G. RUNGER*

SCHOOL OF STATISTICS

UNIVERSITY OF MINNESOTA

Summary

We consider the Box-Cox model of power transformation with special reference to the comparison of two samples. For inferences about a linear model with transformed data we argue that one should behave as if the estimated power transformation were in fact correct, not random. The validity of such inferences is discussed mainly in terms of empirical results.

Some Key Words: Box-Cox model; Data transformation; Normality; Maximum Likelihood.

*Research supported in part by NSF Grant MCS-7904558.

1. Introduction

The scope of normal-theory statistical analysis is widened by allowing preliminary data transformation, based on informal or formal techniques for choice of working scale for the response variable. Among the formal techniques is the Box-Cox analysis, described in Box and Cox (1964), where an estimable transformation is included as one or more parameters of the statistical model. For response variable y and covariable $\underset{\sim}{x}$, Box and Cox considered models such as

$$y(\lambda) = \frac{y^{\lambda} - 1}{\lambda} = \underset{\sim}{\theta}^{T}\underset{\sim}{x} + \sigma e \quad , \tag{1}$$

$y(0) = \log_e(y)$, where λ and the linear model parameters $\underset{\sim}{\theta}$, σ are unknown, and e is hypothetically sampled from a standard normal distribution. Model (1) assumes three things: (i) normality of e, (ii) constant variance of $y(\lambda)$ in repetitions, independent of $\underset{\sim}{x}$, (iii) correctness of the mean form. Our interest will focus only on (i).

The maximum likelihood analysis of data (x_j, y_j), $j=1,\ldots,n$ modelled by (1) may be viewed as first estimating λ by $\hat{\lambda}$, and then estimating $\underset{\sim}{\theta}$ and σ as if $y(\hat{\lambda})$ were normal. But there is an apparent difficulty: under model (1), $\hat{\underset{\sim}{\theta}}$ has an approximate normal distribution with variance in excess of that for known λ. For example, consider the single-sample case with $x_i \equiv 1$ and $\theta = E\{Y(\lambda)\}$. If $\lambda = 0$, then one can easily show that

$$n\,\mathrm{Var}\,(\hat{\theta}) \approx \sigma^2 + \frac{1}{6}\left\{1 + \left(\frac{\theta^2}{\sigma^2}\right)\right\}^2 \quad ; \tag{2}$$

see Hinkley (1975). A comprehensive theoretical and numerical study of this phenomenon has recently been carried out by Bickel and Doksum (1979), who confirm the potentially very large effect on $\mathrm{Var}(\hat{\theta})$ due to estimation of λ .

The purpose of this paper is to describe a conditional analysis of the transformed data which removes the "excess variance" phenomenon from consideration. We believe that in most cases (2) and its generalizations are irrelevant.

To illustrate the essential point, consider two statisticians S_1 and S_2 who are visited by a client C . C has two samples of residual pesticide amounts y in oranges collected in two environments, and he asks S_1 and S_2 for their advice on an appropriate measure of the difference between the two environments. Before proceeding with their data analyses, S_1 and S_2 agree to use the two-sample case of model (1) — barring lack of fit — and to use maximum likelihood estimation. The relative likelihood for λ is given in Figure 1, from which S_1 deduces that the maximum likelihood estimate (m.l.e.) $\hat{\lambda} \approx 0$. Therefore S_1 decides to take logs of the data, and is pleased to find that the transformed data looks normal and has homogeneous variance. At this point S_1 decides to treat $\log_e Y$ as normally distributed, with means μ_A and μ_B in the two sampled populations. Using the data summary

sample A: $\mathrm{mean}(\log_e y) = -5.325$, $\mathrm{s.d.}(\log_e y) = 2.03$, sample size = 21

sample B: $\mathrm{mean}(\log_e y) = -1.820$, $\mathrm{s.d.}(\log_e y) = 2.09$, sample size = 27 ,

S_1 estimates the mean contrast by the 95% confidence interval

$$\mu_A - \mu_B = -5.325 - (-1.820) \pm 2\sqrt{\frac{(2.03)^2}{21} + \frac{(2.09)^2}{27}}$$

$$= -3.51 \pm 1.20 \quad .$$

The report to C reads "The natural log of response was chosen as working scale, since the normality and variance homogeneity assumptions are valid on that scale. Then the contrast between the two environments may be measured by difference in means, for which the 95% confidence interval is -3.51 ± 1.20 ."

S_2, meanwhile, performs the same analysis except that allowance for phenomenon (2) is made in computing the standard error of the mean contrast estimate. This leads to 95% confidence interval

$$\theta_A - \theta_B = -3.51 \pm 3.19 \quad ,$$

and S_2's report says "I cannot be sure of the response scale on which θ_A and θ_B are means for your two populations, but whatever it is their difference has 95% confidence interval -3.51 ± 3.19 . P.S.: the scale is probably close to logarithmic."

Of course one can criticize the choice of contrast measure, where C should have had some say. But given the same choice, the two statisticians give very different answers, and both are approximately correct. We believe that S_1 gives the more useful answer. In Section 2 we discuss the validity of S_1's analysis using theoretical and empirical evidence. Some further general remarks are made in Section 3.

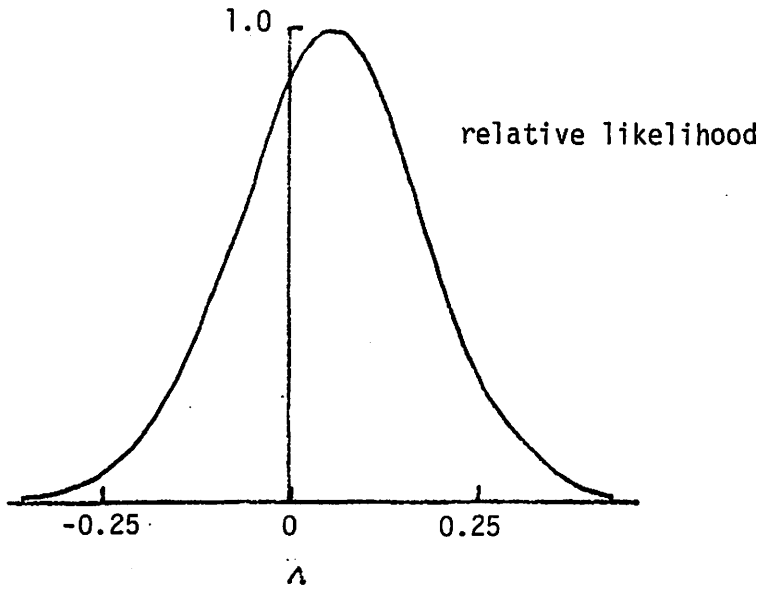

Figure 1. Relative likelihood of transformation power λ for two samples of pesticide data provided by C.

2. Conditional Interpretation of Transformed Data

2.1 Theory

The variance inflation illustrated in (2) may be explained in terms of a simple decomposition for $\hat{\theta} - \theta$. For clarity define

$$\theta(m) = E(\hat{\theta} \mid \hat{\lambda} = m)$$

and write

$$\hat{\theta} - \theta = \hat{\theta} - \theta(\hat{\lambda}) + \theta(\hat{\lambda}) - \theta \quad . \tag{3}$$

Then clearly

$$\mathrm{Var}(\hat{\theta}) = E\{\mathrm{Var}(\hat{\theta} \mid \hat{\lambda})\} + \mathrm{Var}\{\theta(\hat{\lambda}) - \theta\} \ , \tag{4}$$

where the final term is the variance inflation factor due to estimation of λ .

If we wish to condition analysis on the observed value of $\hat{\lambda}$, standard asymptotic theory for maximum likelihood estimation can be used as follows. Write $\psi^T = (\theta^T, \sigma)$ and denote the total information matrix for (ψ^T, λ) and its inverse by

$$I = \begin{bmatrix} I_{\psi\psi} & I_{\psi\lambda} \\ \cdot & I_{\lambda\lambda} \end{bmatrix} \ , \qquad I^{-1} = \begin{bmatrix} I^{\psi\psi} & I^{\psi\lambda} \\ \cdot & I^{\lambda\lambda} \end{bmatrix} \ .$$

Then asymptotically as $n \to \infty$, $\hat{\theta} - \theta$ behaves as $N(0, I^{\psi\psi})$, whereas $\hat{\theta} - \theta$ given $\hat{\lambda} = m$ behaves as $N(-I_{\psi\lambda} I_{\lambda\lambda}^{-1}(m - \lambda) , I_{\psi\psi}^{-1})$; see Cox and Hinkley (1974, §9.3) . That is, the appropriate approximate variance conditional on $\hat{\lambda} = m$ is computed by standard likelihood methods with $\hat{\lambda}$ assumed known equal to m . This was done by S_1 in Section 1, whereas S_2 used

$I^{\psi\psi}$. S_1 was estimating $\theta(\hat{\lambda}_{obs})$, which relates to the θ of model (1) by the approximation

$$\begin{pmatrix}\theta(m)\\ \sigma(m)\end{pmatrix} \approx \begin{pmatrix}\theta\\ \sigma\end{pmatrix} + I_{\psi\lambda} I_{\lambda\lambda}^{-1}(m-\lambda) \quad . \tag{5}$$

In effect S_1 was concluding that the relevant model for analysis is

$$y(\hat{\lambda}_{obs}) = \{\theta(\hat{\lambda}_{obs})\}^T x + \sigma e \quad , \tag{6}$$

where e is standard normal, regardless of which λ makes model (1) true in hypothetical repeated sampling. That is, the hypothetical population defined by model (1) is effectively replaced by a more tangible reference population in which (6) holds — more tangible because the data conform to normality and because the linear model parameter is physically well defined. (Note that properties of normal-theory analysis are not affected by restriction to those normal samples which pass tests of normality based on standarized residuals.)

2.2 Empirical Study

To examine the conditional interpretation (6) empirically, we carried out a small-scale simulation study whose results we summarize here. The first set of results is for the two-sample problem with $\Delta = \theta_1 - \theta_2 = 1$, $\sigma = 1$ and $\lambda = 0$ in model (1). The two sample sizes were $n_1 = n_2 = \frac{1}{2}n = 20$ and 10,000 pairs of samples were generated by standard system algorithm on a CYBER 74. To simulate conditional properties of estimates we grouped samples by 25 interval values of $\hat{\lambda}$ and calculated empirical properties within each group. Figure 2 shows the estimates of $\Delta(\hat{\lambda}) = \theta_1(\hat{\lambda}) - \theta_2(\hat{\lambda})$,

and Figure 3 shows the estimates of both $V(\hat{\lambda}) = \mathrm{Var}(\hat{\Delta}|\hat{\lambda})$ and $\mathrm{Var}\{y(\hat{\lambda})|\hat{\lambda}\}$. The latter is actually the simulation mean of the pooled sample variance estimate $\hat{\sigma}^2(\hat{\lambda})$, and the plotting scale is chosen so that the mean of the estimate of $V(\hat{\lambda})$, namely

$$\hat{V} = \left(\frac{1}{n_1} + \frac{1}{n_2}\right) \hat{\sigma}^2(\hat{\lambda}) \quad , \tag{7}$$

is on the same scale as $V(\hat{\lambda})$. The agreement is clearly good. Notice in Figure 1 that the approximation to $\Delta(m)$ based on (5) is poor, presumably because it is only a local approximation.

Standard 95% confidence limits for Δ were calculated assuming (6), that is the limits were

$$\hat{L},\ \hat{U} = \hat{\theta}_1 - \hat{\theta}_2 \pm 2\hat{V}^{\frac{1}{2}} \quad . \tag{8}$$

with $\hat{V}$ given by (7). The empirical coverage probabilities for $(\hat{L},\hat{U})$ are shown in Figure 4 (see Appendix 1 for relevant methodology). Agreement with nominal 95% is good.

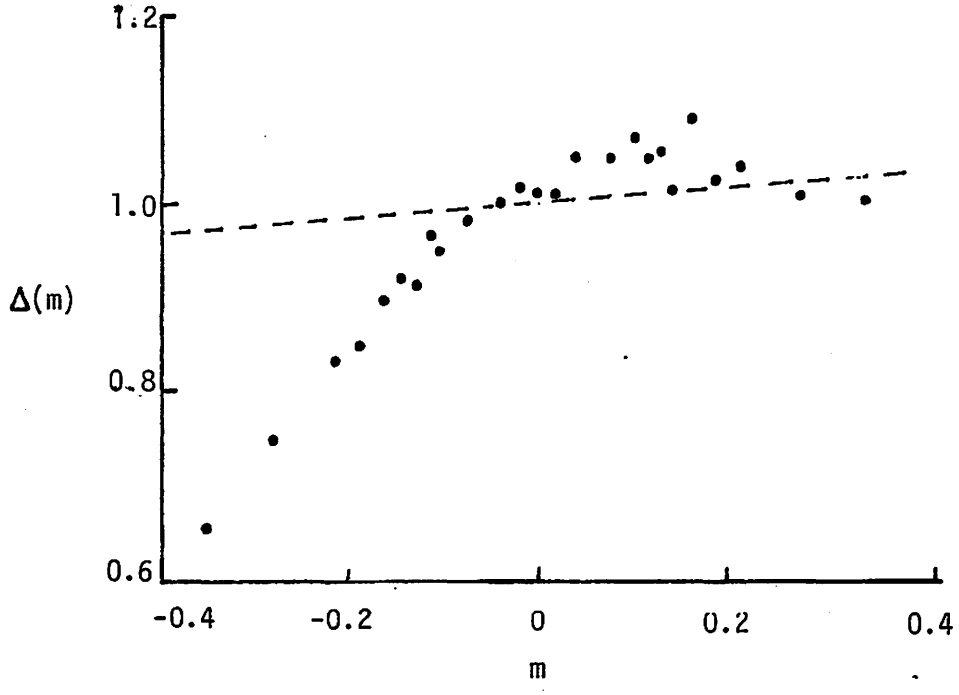

Figure 2. Conditional empirical mean of $\hat{\Delta} = \hat{\theta}_1 - \hat{\theta}_2$ in two-sample case of model (1) when $\Delta = 1$; $\sigma = 1$, $\lambda = 0$, $n_1 = n_2 = 20$. Dotted line derived from equation (5).

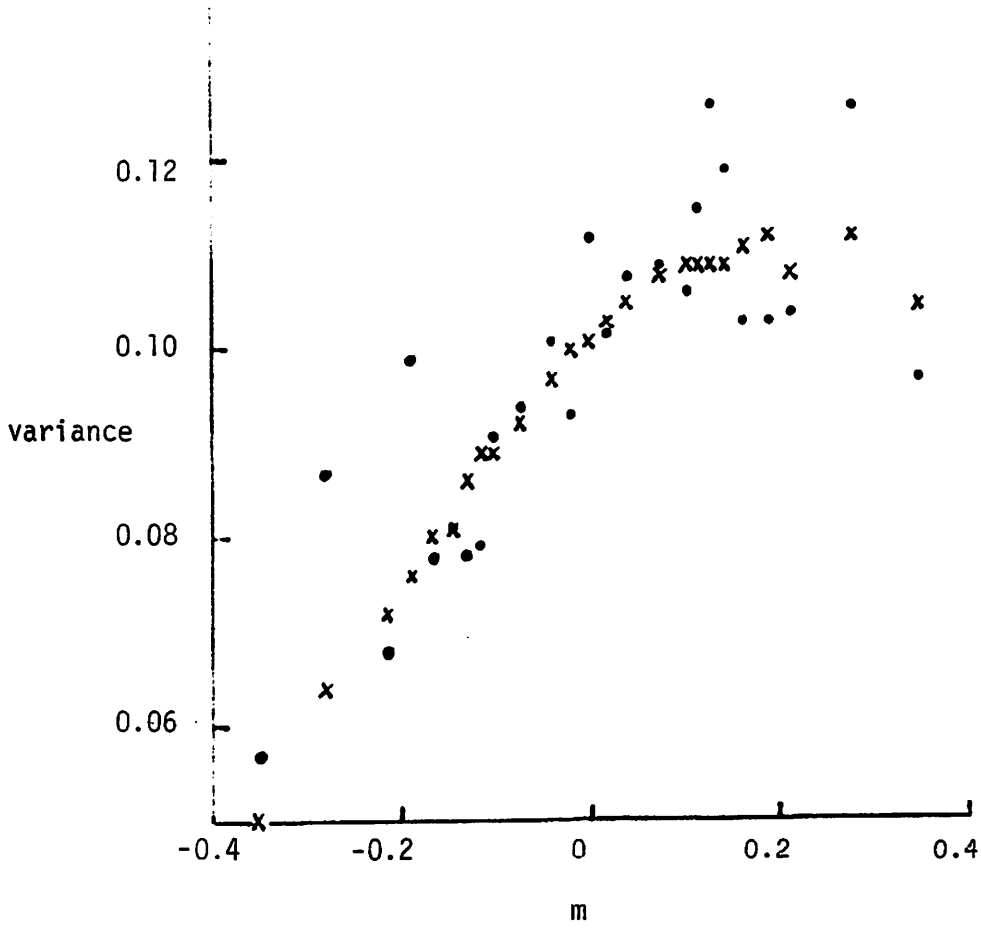

Figure 3. Conditional empirical variance of $\hat{\Delta}$ and mean of estimated variance $\hat{V}$ in the two-sample case of model (1) with $\Delta = 1$, $\sigma = 1$, $\lambda = 0$, $n_1 = n_2 = 20$. Note that $\hat{V} = \hat{\sigma}^2(\hat{\lambda})/10$.
Key: • V(m) ; × mean $\hat{V}$

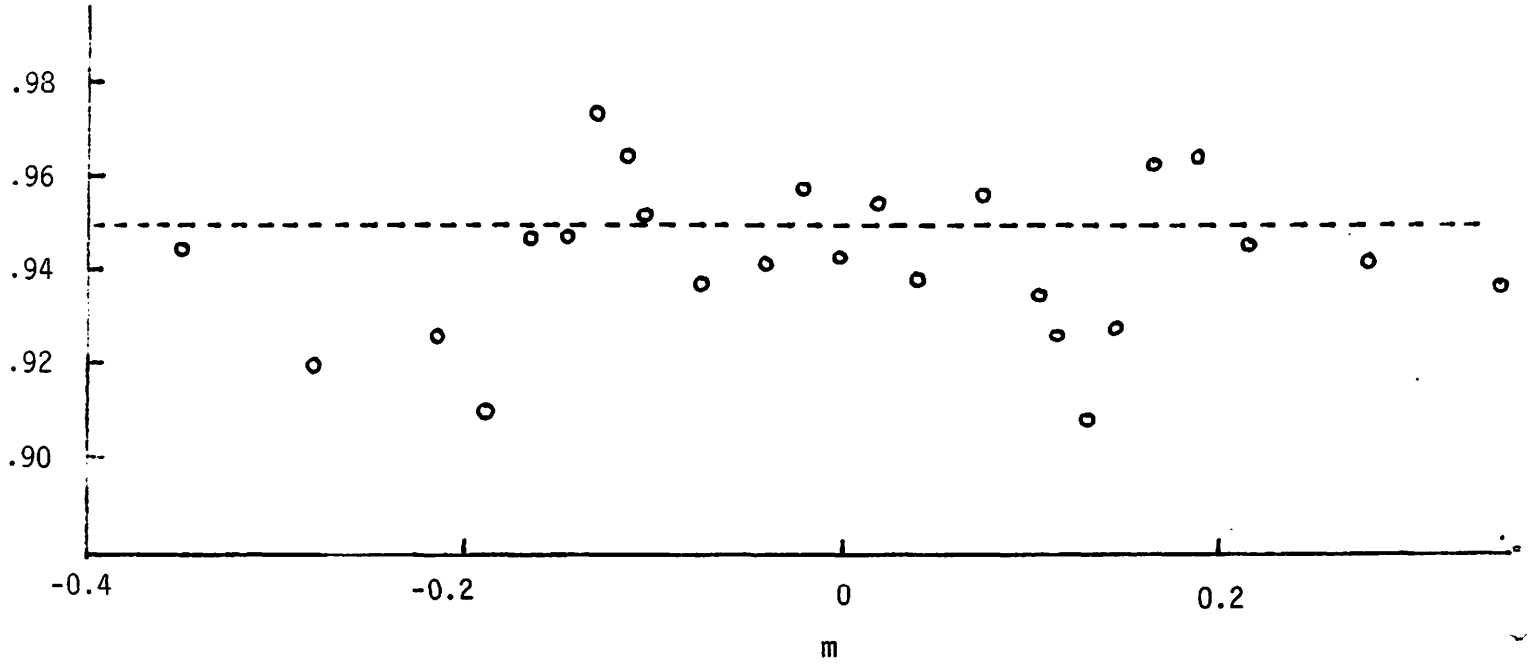

Figure 4. Empirical coverage frequencies of normal 95% confidence intervals (8) for Δ .

In practice it would be quite usual to restrict λ to a small discrete set, for example $\{-2, -1, -\frac{1}{2}, 0, \frac{1}{2}, 1, 2\}$. The following results were obtained from small simulations with this restriction, when the "true" λ is -1, 0 or +1 . The specific models simulated were again two-sample models with characteristics as given in Table 1.

Table 1. Cases of model (1) used in simulation

λ	θ_1	θ_2	σ	$n_1 = n_2$	# pairs of samples
-1	-2	$-2\frac{1}{6}$	$\frac{1}{6}$	20	1000
0	$\frac{1}{4}$	0	$\frac{1}{4}$	20	1000
+1	$\frac{1}{6}$	0	$\frac{1}{6}$	20	1000

The outcomes of these simulations are summarized in Table 2, including empirical frequencies for $\hat{\lambda}$. For each characteristic the results are given in the vertical order $\lambda = -1, 0, 1$. The entries for $\lambda = 1$, $\hat{\lambda} = \frac{1}{2}$ are a little peculiar — here the variance $V(\hat{\lambda})$ was unusually large, 30% larger than the mean of $\hat{V}$. Overall the results look very good for validity of confidence intervals based on (6).

Table 2. Monte Carlo Estimates of Conditional Properties of $\hat{\Delta}$ for Cases in Table 1. (Blank indicates values of $\hat{\lambda}$ not allowed.)

m	-2	-1	$-\frac{1}{2}$	0	$\frac{1}{2}$	1	2
$pr(\hat{\lambda} = m)$	.227	.384	.193	.122	.074	-	-
	-	.068	.241	.394	.239	.058	-
	-	-	.071	.118	.199	.385	.227
$\Delta(m) = E(\hat{\theta}_1 - \hat{\theta}_2 \mid \hat{\lambda}=m)$	.166	.171	.154	.155	.144	-	-
	-	.191	.241	.250	.271	.256	-
	-	-	.134	.162	.161	.169	.166
Conditional coverage of "95%" intervals for $\Delta(m)$	.925	.932	.938	.951	.946	-	-
	-	.926	.954	.944	.937	.948	-
	-	-	.944	.983	.894	.948	.956
Conditional coverage of "90%" intervals for $\Delta(m)$	.881	.904	.870	.910	.824	-	-
	-	.897	.913	.901	.870	.897	-
	-	-	.915	.941	.854	.891	.907

2.3 Normality of Transformed Data

Even if we regard the results of the simulations as convincing evidence that S_1 gave a valid analysis, it is relevant to question the normality of $y(\hat{\lambda})$, since the efficiency of S_1's analysis depends on it. One might take the view that if the $y(\hat{\lambda})$ look like normal data, then we are justified in assuming normality and rejecting the more abstract model (1). But there will be situations where model (1) holds — what of $y(\hat{\lambda})$ then?

Our evidence suggests that the distribution of $y(\hat{\lambda})$ conditional on $\hat{\lambda}$ will be close to normal, deviating in the direction of shorter tails. This is to be expected because estimation of λ is associated with shrinking of relatively large residuals: a single-sample configuration such as

will be transformed to reduce the last, underlined gap.

Our very limited empirical evidence comes from a single-sample Monte Carlo experiment with 5000 samples of the case $n = 20$, $\theta = 0$, $\sigma = 3$ and $\lambda = 0$. Table 3 gives the estimated conditional standardized cumulants of $y(\hat{\lambda})$ for 13 interval values of $\hat{\lambda}$. These results show significant evidence of small deviations from normality in conditional (and unconditional) distributions of $y(\hat{\lambda})$. One can assess the approximate impact of these deviations on the efficiency of least-squares methods using the approximations given by Cox and Hinkley (1968, §4). For example, if standardized third and fourth cumulants are respectively 0.25 and 0.4, then the approximate efficiency of least squares is 92%. We stress that

this figure is based on assuming the truth of model (1), which is hypothetical, whereas we would argue that one should model the transformed data by (6).

Within the hypothetical framework of model (1), the distribution of $y(\hat{\lambda})$ can be studied theoretically. Some thoughts on this are outlined in Appendix 2.

Table 3. Estimates of Standardized Conditional Cumulants of $y(\hat{\lambda})$ in the Single-Sample Case $\theta = 0$, $\sigma = 0$, $\lambda = 0$, $n = 20$ (5000 samples).

m	-0.75	-0.5	-0.3	-0.2	-0.1	-.05	0	.05	.1	.2	.3	.5	.75
$pr(\hat{\lambda} = m)$	.0002*	.0036	.0186	.0750	.1554	.1666	.1958	.1590	.1392	.0662	.0188	.0014	.0002*
1st cumulant μ	-.24	-.51	-.80	-.61	-.45	-.22	.02	.25	.40	.80	.64	.34	.10
2nd cumulant σ^2	0.56	1.95	3.93	5.40	7.30	8.54	8.87	8.62	7.21	5.92	3.40	1.98	0.50
3rd cumulant γ_1	0.74	0.30	0.26	0.13	0.13	0.04	0.00	-0.04	-0.14	-0.14	-0.30	0.74	0.02
4th cumulant γ_2	-	+0.62	-0.25	-0.41	-0.40	-0.46	-0.39	-0.41	-0.43	-0.27	0.01	2.95	0.59

*only one sample

3. Further Remarks

In the (real) data analysis problem of Section 1, the choice of comparative measure might not be thought appropriate. For example, one might wish to compare means or medians on the original scale. With medians one can transform back from analyses on the $y(\hat{\lambda})$ scale, where mean=median, and the variance inflation effect can be ignored; see Ruppert and Carroll (1980). Comparison of means would be somewhat more difficult.

The situation that we have considered is that of a single analysis, as opposed to a series of analyses with similar data. If a series of analyses give an overall estimate $\tilde{\lambda}$, then presumably the individual analyses would all be carried out using $y(\tilde{\lambda})$ and <u>not</u> the separate transformations $y(\hat{\lambda}_1)$, $y(\hat{\lambda}_2)$,... . In that sense, there is no operational meaning to a series of $\theta(\hat{\lambda})$ estimates with $\hat{\lambda}$ varying, in a manner determined by model (1).

One standard alternative to $y(\lambda)$ is the transformed variable $z(\lambda) = y(\lambda)/\dot{y}^{\lambda-1}$, where $\dot{y}$ is the geometric mean. We found $z(\hat{\lambda})$ to be quite badly behaved, in the senses that (i) its conditional mean varied much more than that of $y(\hat{\lambda})$ and (ii) the conditional standard error of linear model parameters was poorly estimated using residual variance estimates.

References

Bickel, P. J. and Doksum, K. A. (1979). An analysis of transformations revisited. Univ. of Calif., Berkeley, Dept. of Statistics Report.

Box, G. E. P. and Cox, D. R. (1964). An analysis of transformations (with discussion). J. R. Statist. Soc., B, 26, 211-252.

Hinkley, D. V. (1975). On power transformations to symmetry. Biometrika, 62, 101-112.

Carroll, R. J. and Ruppert, D. (1980). On prediction and the power transformation family. Dept. of Stat., UNC Chapel Hill, Inst. of Stat. Mimeo Series, #1264.

Reeds, J. A. (1976). On the definition of von Mises functionals. Ph.D. thesis, Harvard University.

Appendix 1: Computation of Coverage Frequencies

When simulating the coverage probabilities (8), an initial small simulation was used to estimate $\Delta(m) = \theta_1(m) - \theta_2(m)$. Denote this estimate by $\Delta_1(m)$. Then the main large simulation gave relative frequencies of the event

$$\hat{L} \leq \Delta_1(m) \leq \hat{U}$$

and a second, very precise, estimate $\Delta_2(m)$ of $\Delta(m)$. Denote the estimate of $\Delta(m)$ from a single pair of samples by $\hat{\Delta}$. Then we assume that, conditional on $\hat{\lambda} = m$,

$$\frac{\hat{\Delta} - \Delta_2(m)}{\sqrt{V(m)}}$$

is exactly normal. Therefore the relative frequence of $\hat{L} \leq \Delta_1(m) \leq \hat{U}$ estimates

$$\Phi\left(2 + \frac{\Delta_2(m) - \Delta_1(m)}{\sqrt{V(m)}}\right) - \Phi\left(-2 + \frac{\Delta_2(m) - \Delta_1(m)}{\sqrt{V(m)}}\right)$$

$$\doteq 2\Phi(-2) - \frac{\{\Delta_2(m) - \Delta_1(m)\}^2}{V(m)} \; 2\phi(2) \quad , \tag{A1}$$

rather than $2\Phi(-2)$. Therefore we adjust the relative frequency of $\hat{L} \leq \Delta_1(m) \leq \hat{U}$ by adding the second term on the right of (A1). This adjustment is employed in Figure 3 and Table 2.

Appendix 2: Empirical Distribution of Transformed Data

Properties of the transformed variables $Y(\hat{\lambda})$ can be determined approximately by use of a simple functional representation that we describe here. For simplicity we consider the single-sample situation.

Let the empirical distribution function of sample $Y_1,\ldots,Y_n$ be

$$\hat{F}_n(y) = \frac{1}{n} \sum_{j=1}^{n} I(y - Y_j) \quad ,$$

where $I(u) = 1$ $(u \geq 0)$, $0(u < 0)$. Then, with $\hat{\lambda} = \lambda(\hat{F}_n)$, $\hat{\mu} = \mu(\hat{F}_n)$ and $\hat{\sigma} = \sigma(\hat{F}_n)$, let

$$G^*(z;\ \hat{F}_n) = \frac{1}{n} \sum_{j=1}^{n} I\left(z - \frac{Y_j(\hat{\lambda}) - \hat{\mu}}{\hat{\sigma}}\right) \quad ,$$

the empirical distribution of the standardized transformed sample. The corresponding population functional, to which $G^*(z;\ \hat{F}_n)$ converges as $n \to \infty$, is easily seen to be

$$G^*(z;\ F) = \int I\left\{\frac{\log_e[1 + \lambda(F)\{\mu(F) + z\ \sigma(F)\}]}{\lambda(F)} - y\right\} dF(y) \quad .$$

Here F is the c.d.f. of Y determined by model (1), and of course evaluation of $G^*(z;\ F)$ gives the standard normal c.d.f.

What is of interest is the behavior of $G^*(z;\hat{F}_n)$ relative to $G^*(z;F)$, which involves the Taylor series expansion

$$G^*(z;\hat{F}_n) = G^*(z;F) + \int \dot{G}^*(z;F;u)\,d(\hat{F}_n - F)(u)$$

$$+ \frac{1}{2}\iint \ddot{G}^*(z;F;u,v)\,d(\hat{F}_n - F)(u)\,d(\hat{F}_n - F)(v) + \ldots$$

with $\dot{G}^*$, $\ddot{G}^*$, etc. the successive von Mises derivatives of G (Reeds, 1976). For example, one can compute series expansions for the mean of $G^*(z;\hat{F}_n)$ or its moments. In order to determine the derivatives of G^* , simple chain-rule arguments and the derivatives $\dot{\lambda}(F;u)$, $\dot{\mu}(F;u)$, $\dot{\sigma}(F;u)$, etc. are used.

In the context of Section 2, particular interest focusses on the behavior of $G^*(z;\hat{F}_n)$ conditional on $\lambda(\hat{F}_n)$, where $\lambda(\hat{F}_n)$ has a Taylor series expansion also.