The Annals of Statistics
1992, Vol. 20, No. 3, 1485–1500

BOUNDS ON AREs OF TESTS FOLLOWING BOX–COX TRANSFORMATIONS

BY HANFENG CHEN[1] AND WEI-YIN LOH[2]

Bowling Green State University and University of Wisconsin, Madison

Bounds on the asymptotic relative efficiency (ARE) of the Box–Cox transformed two-sample t-test to the ordinary t-test are obtained under local alternatives. It is shown that the ARE is at least 1 for location-shift models. In the case of scale-shift models, a similar bound applies provided the limiting value of the estimated power transformation is greater than 1. If instead the Box–Cox transformed t-test is compared against the ordinary t-test applied to the log-transformed data, then the ARE is bounded below by 1 for all scale-shift models, regardless of the limiting value of the power transformation. The results extend naturally to the multisample F-test. A small simulation study to evaluate the validity of the asymptotic results for finite-sample sizes is also reported.

1. Introduction. Standard linear model analyses of data usually assume that the observations are independent and normally distributed, with constant variance and with expectations specified by a model linear in its set of parameters. In an oft-cited paper, Box and Cox (1964) suggested that these assumptions can be weakened so that they hold after some suitable transformation of the observations. The power family of transformations is one particular family which is amenable to analysis and appears to be quite fruitful in applications. For example, Box and Cox applied their method to data from a 3×4 factorial experiment and showed that after a reciprocal transformation, the sensitivity of their analysis of variance F-test was "increased almost threefold" [Box and Cox (1964), page 222]. Box, Hunter and Hunter (1978), page 238, interpret this as "equivalent to increasing the size of the experiment by a factor of nearly three." Other examples can be found in Box and Cox (1964) and Atkinson (1985).

Theoretical studies on several aspects of the Box–Cox method have been reported in the literature. Hinkley (1975) and Hernandez and Johnson (1980) investigated the asymptotic properties of the parameter estimates for the one-sample problem. Bickel and Doksum (1981) examined the behavior of the asymptotic variances of parameter estimates for regression and analysis of variance situations. Doksum and Wong (1983) showed that in certain circumstances some tests have asymptotically the correct level and asymptotically the

Received October 1990; revised September 1991.
[1]Research partially supported by a Bowling Green State University Faculty Research Committee Basic Grant.
[2]Research partially supported by NSF Grant DMS-88-03271, ARO Grant DAAL03-91-G-0111 and by funds from the University of Wisconsin Graduate School.
AMS 1980 *subject classifications*. Primary 62G20; secondary 62G10.
Key words and phrases. Asymptotic efficiency, location shift, scale shift, t-test.

same local power as if the value of the transformation parameter were known. Hinkley and Runger (1984) approached the problem from a conditional viewpoint [see also Doksum (1987)], and Carroll (1980) looked at the robust choice of transformations.

The present paper was motivated by Box and Cox's suggestion that a mere transformation can achieve the equivalent of a threefold increase in sample size for a hypothesis test. The results of an analytical and numerical study are reported here on the following simpler but common practical problem. Suppose we have two sets of observations from distributions with support on the positive half-line, and the distributions are assumed to differ by a location shift at most. A natural test of the null hypothesis of no shift is the two-sample t-test, which is optimal when the distributions are normal with constant variance. Because the observations cannot take negative values, however, the distributions can only be approximately normal at best. It is reasonable, therefore, to ask if a transformation chosen to improve the normality of the transformed data would enhance the sensitivity of the two-sample t-test. Note that since the data are already assumed to be homoscedastic from the start, any nonlinear transformation will necessarily cause them to be heteroscedastic after transformation, if the null hypothesis is false. To avoid severe heteroscedasticity, the transformation must therefore be chosen to balance the goals of normality and homoscedasticity as much as possible. Interestingly, if the null hypothesis is false, the distributions after transformation could differ by more than just a location shift, and the t-test, which is meant to detect shifts in location, may instead lose its sensitivity after transformation. On the other hand, if the null hypothesis is true, the two distributions would remain identical after transformation, and hence the significance level of the t-test applied to the transformed data will still be approximately valid, especially with large samples.

It turns out that the transformation to approximate normality and homoscedasticity more than offsets any perturbations in the shapes of the distributions, because the transformed t-test is always more powerful asymptotically than the untransformed t-test. Thus Box and Cox's expectation holds quite generally. A parallel analysis for scale-shift models is also presented. First we need some notation.

Let $(X_1, \dots, X_m)$ and $(V_1, \dots, V_n)$ be independent copies of a positive random variable X, and let $\tau \geq 0$ and $N = m + n$. Let the variables $Y_1, \dots, Y_n$ be defined by

$$Y_j = V_j + \tau N^{-1/2}, \qquad j = 1, \dots, n, \tag{1}$$

in which case we say that the X's and Y's follow a *location-shift* model. We say they follow a *scale-shift* model if the Y's are defined instead by

$$Y_j = (1 + \tau N^{-1/2}) V_j, \qquad j = 1, \dots, n. \tag{2}$$

Consider testing the hypotheses

$$H_0\colon \tau = 0 \quad \text{versus} \quad H_1\colon \tau > 0 \tag{3}$$

given that only the X's and Y's are observed and assuming that $m/n \to k$ as $m, n \to \infty$, for some $k \in (0, \infty)$. The analyses of models (1) and (2) can be carried out simultaneously if we embed them in the more general model

$$Y_j = (1 + c_N)^r (V_j + c_N u), \qquad j = 1, \dots, n, \tag{4}$$

where $r \geq 0$, $u \geq 0$, $r + u > 0$ and $c_N = \tau N^{-1/2}$. Clearly, the location-shift and scale-shift models correspond to the cases ($r = 0$, $u = 1$) and ($r = 1$, $u = 0$), respectively.

Given a number λ, define

$$X_i(\lambda) = \begin{cases} (X_i^\lambda - 1)/\lambda, & \text{if } \lambda \neq 0, \\ \log X_i, & \text{if } \lambda = 0, \end{cases}$$

with a similar definition for $Y_j(\lambda)$. The value of λ may be estimated from the data by maximum likelihood or Bayes theory. We consider only maximum likelihood estimation here and denote the estimate by $\hat{\lambda}$. Note that because the data have support on the positive half-line, there does not in general exist a value of λ that makes the transformed data exactly normal, except when they are log-normally distributed to begin with.

We consider the question of the relative efficiency of tests of (3) with and without transformation. Doksum and Wong (1983) considered a related question under the following assumptions. Suppose there exists a λ_1 such that $X_i(\lambda_1) = a + \tau\varepsilon_i$ and $Y_j(\lambda_1) = b + \tau\varepsilon_j$, where the ε_i's have a common distribution symmetric about 0, and suppose that the maximum likelihood estimate of λ_1 exists and is consistent. Doksum and Wong gave an expression for the ARE of the t-test of the hypothesis H_0: $a = b$ based on the transformed data to that based on the untransformed data. They therefore assumed that the populations differ by a location shift in the (unknown) *transformed* scale. Another situation they considered was that the populations differ by a scale shift in the *untransformed* scale.

We assume that the populations differ only by a location shift or a scale shift in the untransformed scale because this is the standard framework used in the literature on rank tests. Under these models, we establish sufficient conditions for $\hat{\lambda} \to \lambda_0$ a.s. and for the asymptotic normality of $N^{1/2}(\hat{\lambda} - \lambda_0)$ in Section 2. General expressions for the asymptotic relative efficiencies (AREs) of the t-tests with and without transformation are derived in Section 3. Because of the different model assumptions, our AREs in the location-shift case are different from those obtained by Doksum and Wong (1983), equation (3.3). However, the AREs in the scale-shift case are the same. Sharp bounds for the AREs are obtained for the two-sample and multisample problems. The bounds show, for example, that the Box–Cox transformed tests are always more efficient asymptotically than the untransformed tests under location-shift alternatives.

A similar conclusion does not hold in the case of scale shift, where it turns out that the ARE is bounded below by 1 only if $\lambda_0 \geq 1$. If one believes in a scale-shift model, however, it may be more reasonable to apply the t-test to the log-transformed data instead of the original data, because the log transformation changes a scale-shift model to a location-shift one. We prove that the asymptotic efficiency of the Box–Cox t-test is also always at least as good as that of the log-transformed t-test for this situation. Numerical values to demonstrate the size of the AREs are reported for some distributions in Section 4.

The results of a simulation experiment to study the accuracy of the asymptotic predictions for fixed sample sizes are reported in Section 5, and Section 6 presents generalizations of the theoretical results to the multisample F-test.

2. Asymptotic properties. Let $\bar{X}(\lambda) = m^{-1}\sum_{i=1}^{m} X_i(\lambda)$, with a similar definition for $\bar{Y}(\lambda)$. Let $s_N^2(\lambda)$ be the pooled estimate of variance of the transformed data and let $t_N(\lambda)$ be the t-statistic based on them, that is,

$$s_N^2(\lambda) = \frac{1}{N-2}\left[\sum_{i=1}^{m}\{X_i(\lambda) - \bar{X}(\lambda)\}^2 + \sum_{j=1}^{n}\{Y_j(\lambda) - \bar{Y}(\lambda)\}^2\right],$$

$$t_N(\lambda) = (mn/N)^{1/2} s_N^{-1}(\lambda)\{\bar{Y}(\lambda) - \bar{X}(\lambda)\}.$$

The maximum likelihood estimate $\hat{\lambda}$ is the minimizer of the function

$$J_N(\lambda) = s_N^2(\lambda)/\dot{z}^{2\lambda},$$

where $\log(\dot{z}) = N^{-1}\{\sum_{i=1}^{m}\log(X_i) + \sum_{j=1}^{n}\log(Y_j)\}$. Clearly, if $EX^{2\lambda}$ and $E\log X$ are finite, then $J_N(\lambda) \to J_0(\lambda)$ a.s. under H_0, where

$$J_0(\lambda) = \sigma^2(\lambda)\exp(-2\lambda\eta),$$

$\sigma^2(\lambda) = \mathrm{Var}\{X(\lambda)\}$ and $\eta = E\log X$. The minimizer of $J_0(\lambda)$ will be denoted by λ_0.

We assume there is a finite closed interval $[a, b]$ over which the following conditions are satisfied:

$$EX^{2\lambda} < \infty, \qquad \lambda \in [a, b], \tag{5}$$

$$E|\log X| < \infty, \tag{6}$$

$$E(X^{\lambda}\log X)^2 < \infty, \qquad \lambda \in [a, b], \tag{7}$$

$$\hat{\lambda}, \lambda_0 \in (a, b), \tag{8}$$

$$\lambda_0 \text{ is unique on } [a, b], \tag{9}$$

$$EX(\lambda) < \infty, \qquad \lambda \in [a-1, b]. \tag{10}$$

The first and second derivatives of a function f will be denoted by f' and f'', respectively.

LEMMA 1. *Assume conditions* (4)–(7). *Then under* H_0 *as well as* H_1, $J_N(\lambda) \to J_0(\lambda)$ *and* $J_N''(\lambda) \to J_0''(\lambda)$ *a.s. uniformly in* $\lambda \in [a, b]$.

PROOF. The basic idea depends on Rubin's theorem [Rubin (1956), Theorem 1]. It suffices to prove the case for $J_N(\lambda)$ since that for $J_N''(\lambda)$ is similar. Let

$$f(\lambda, x) = \begin{cases} (x^\lambda - 1)/\lambda, & \lambda \neq 0, \\ \log(x), & \lambda = 0. \end{cases}$$

For each fixed value of x, $f(\lambda, x)$ is a continuous monotone function of λ. Define $g_1(x) = \max\{|f(a, x)|, |f(b, x)|\}$. Then $|f(\lambda, x)| \le g_1(x)$, $\lambda \in [a, b]$. Let $S_i = [-i, i]$. For each i, $f(\lambda, x)$ is equicontinuous in λ for $x \in S_i$. By Rubin's theorem, we see that with probability 1, $\bar{X}(\lambda) = m^{-1}\sum_{k=1}^{m} f(\lambda, X_k) \to EX(\lambda)$ uniformly for $\lambda \in [a, b]$ as $m \to \infty$.

Next define $h(c, \lambda, x) = \lambda^{-1}\{[(1+c)^d(x + cu)]^\lambda - 1\}$, $c \in [0, 1]$, $\lambda \in [a, b]$ and

$$g_2(x) = \max\{|h(0, a, x)|, |h(0, b, x)|, |h(1, a, x)|, |h(1, b, x)|\}.$$

Clearly, $|h(c, \lambda, x)| \le g_2(x)$ for $c \in [0, 1]$ and $\lambda \in [a, b]$. Another application of Rubin's theorem shows that with probability 1, $n^{-1}\sum_{k=1}^{n} h(c, \lambda, V_k) \to Eh(c, \lambda, X) = EY_1(\lambda)$ uniformly in $c \in [0, 1]$ and $\lambda \in [a, b]$ as $n \to \infty$. Because $\lim_{c \to 0} EY_1(\lambda) = EX(\lambda)$, we conclude that under H_0 and H_1, $\bar{Y}(\lambda) = n^{-1}\sum_{k=1}^{n} h(c, \lambda, V_k) \to EX(\lambda)$ a.s. uniformly for $\lambda \in [a, b]$ as $N \to \infty$.

It is similarly proved that with probability 1 and uniformly in $\lambda \in [a, b]$,

$$\frac{1}{m}\sum_{k=1}^{m} X_k^2(\lambda) \to EX^2(\lambda), \qquad \frac{1}{n}\sum_{k=1}^{n} Y_k^2(\lambda) \to EX^2(\lambda),$$

$$s_N^2(\lambda) \to \sigma^2(\lambda), \qquad \dot{z}^{2\lambda} \to \exp\{2\lambda\eta\}.$$

Therefore with probability 1, $J_N(\lambda) \to J_0(\lambda)$ uniformly for $\lambda \in [a, b]$. □

THEOREM 1. *Assume conditions* (4)–(9). *Then under* H_0 *and* H_1, (i) $\hat{\lambda}$ *converges almost surely to* λ_0 *and* (ii) $N^{1/2}(\hat{\lambda} - \lambda_0)$ *has a normal limit distribution.*

PROOF. By Lemma 1, $J_N(\lambda) \to J_0(\lambda)$ a.s. uniformly on the interval $[a, b]$. Condition (9) and the continuity of $J_0(\lambda)$ imply part (i) of the theorem. To prove part (ii), apply the Taylor expansion to $J_N'(\hat{\lambda})$ at λ_0 to obtain

$$0 = J_N'(\hat{\lambda}) = J_N'(\lambda_0) + J_N''(\lambda^*)(\hat{\lambda} - \lambda_0),$$

where $|\lambda^* - \lambda_0| \le |\hat{\lambda} - \lambda_0|$. Hence $N^{1/2}(\hat{\lambda} - \lambda_0) = -[J_N''(\lambda_*)]^{-1}N^{1/2}J_N'(\lambda_0)$. By part (i) of the theorem and Lemma 1, $J_N''(\lambda^*) \to J_0''(\lambda_0)$ a.s. Since $N^{1/2}J_N'(\lambda_0)$ has a normal limit distribution, $N^{1/2}(\hat{\lambda} - \lambda_0)$ also converges to a normal limit. □

THEOREM 2. *Assume conditions* (4)–(9). *Then under* H_0 *and* H_1, $t_N(\hat{\lambda}) - t_N(\lambda_0) = o_P(1)$.

PROOF. Write

$$t_N(\hat{\lambda}) - t_N(\lambda_0) = (mn/N)^{1/2}(\hat{\lambda} - \lambda_0)(mn/N)^{-1/2}t'_N(\lambda^*),$$

where $|\lambda^* - \lambda_0| \le |\hat{\lambda} - \lambda_0|$ and

$$(mn/N)^{-1/2}t'_N(\lambda) = s_N^{-2}(\lambda)\big[s_N(\lambda)\{\bar{Y}'(\lambda) - \bar{X}'(\lambda)\} - \{\bar{Y}(\lambda) - \bar{X}(\lambda)\}s'_N(\lambda)\big].$$

By Theorem 1, $(mn/N)^{1/2}(\hat{\lambda} - \lambda_0) = O_P(1)$, so it suffices to show that $(mn/N)^{-1/2}t'_N(\lambda^*)$ converges to 0. It may be seen from the proof of Lemma 1 that the sample mean and variance of the transformed data, together with their derivatives, converge a.s. uniformly in λ. Further, according to Theorem 1, $\lambda^* \to \lambda_0$ a.s. Therefore $\bar{Y}'(\lambda^*) - \bar{X}'(\lambda^*) \to 0$ a.s. and $\bar{Y}(\lambda^*) - \bar{X}(\lambda^*) \to 0$ a.s. Hence $(mn/N)^{-1/2}t'_N(\lambda^*) \to 0$ a.s. and so $t_N(\hat{\lambda}) - t_N(\lambda_0) = o_P(1)$. □

Define

$$\xi_\lambda = \tau k^{1/2}(k+1)^{-1}(uEX^{\lambda-1} + rEX^\lambda)[\mathrm{Var}\{X(\lambda)\}]^{-1/2}$$

$$(11) \quad = \begin{cases} \tau k^{1/2}(k+1)^{-1}|\lambda|(uEX^{\lambda-1} + rEX^\lambda)\{\mathrm{Var}(X^\lambda)\}^{-1/2}, & \text{if } \lambda \ne 0, \\ \tau k^{1/2}(k+1)^{-1}(uEX^{-1} + r)\{\mathrm{Var}(\log X)\}^{-1/2}, & \text{if } \lambda = 0, \end{cases}$$

and let "$\to_D$" denote convergence in distribution.

THEOREM 3. *Suppose that conditions* (4)–(7) *and* (10) *hold. Then under* H_0 *as well as* H_1, $t_N(\lambda) \to_D N(\xi_\lambda, 1)$ *and* $t_N^2(\lambda) \to_D \chi_1^2(\delta_\lambda)$, *where* $\delta_\lambda = \xi_\lambda^2$ *is the noncentrality parameter of the noncentral* χ^2 *distribution with one degree of freedom.*

PROOF. Write

$$t_N(\lambda) = \tilde{t}_N(\lambda) + \delta_N(\lambda),$$

$$\tilde{t}_N(\lambda) = (mn/N)^{1/2}\{\bar{Y}(\lambda) - \bar{X}(\lambda) - [EY(\lambda) - EX(\lambda)]\}/s_N(\lambda),$$

$$\delta_N(\lambda) = (mn/N)^{1/2}\{EY(\lambda) - EX(\lambda)\}/s_N(\lambda).$$

It follows that $\tilde{t}_N(\lambda) \to_D N(0,1)$ and $\delta_N(\lambda) \to \xi_\lambda$ a.s. Hence $t_N(\lambda) \to_D N(\xi_\lambda, 1)$. □

COROLLARY 1. *Assume conditions* (4), (5) *and* (10) *hold. Then* $t_N(\hat{\lambda}) \to_D N(\xi_{\lambda_0}, 1)$ *and* $t_N^2(\hat{\lambda}) \to_D \chi_1^2(\delta_{\lambda_0})$.

3. Asymptotic relative efficiency (ARE). The following notation is introduced for convenience. Denote by t_{BC} the Box–Cox t-test [i.e., the test based on $t_N(\hat{\lambda})$] and define

$$(12)\qquad \begin{aligned} A_{\lambda_0} &\equiv \begin{cases} \sigma|\lambda_0|EX^{\lambda_0-1}\{\mathrm{Var}(X^{\lambda_0})\}^{-1/2}, & \text{if } \lambda_0 \neq 0,\\ \sigma EX^{-1}\{\mathrm{Var}(\log X)\}^{-1/2}, & \text{if } \lambda_0 = 0,\end{cases}\\ B_{\lambda_0} &\equiv \begin{cases} \sigma|\lambda_0|(EX^{\lambda_0}/EX)\{\mathrm{Var}(X^{\lambda_0})\}^{-1/2}, & \text{if } \lambda_0 \neq 0,\\ \sigma(EX)^{-1}\{\mathrm{Var}(\log X)\}^{-1/2}, & \text{if } \lambda_0 = 0,\end{cases}\end{aligned}$$

where $\sigma^2 = \mathrm{Var}(X)$.

THEOREM 4. *Assume conditions* (4)–(10). *Then the ARE of t_{BC} to the ordinary t-test is $ARE(t_{\mathrm{BC}}, t) = (pA_{\lambda_0} + qB_{\lambda_0})^2$, where $p = u/|rEX + u|$ and $q = rEX/|rEX + u|$.*

PROOF. Against a one-sided alternative, t_{BC} rejects the null hypothesis whenever $t_N(\hat{\lambda})$ is large. Given τ and N, let $\beta(c_N|\phi, M)$ be the power of test ϕ at the local alternative defined in (3) and (4) based on sample size M. Define N' to be the solution of the equation $\beta(c_N|t_{\mathrm{BC}}, N) = \beta(c_N|t, N')$ and let $\gamma = \mathrm{ARE}(t_{\mathrm{BC}}, t) \equiv \lim_{N\to\infty} N'/N$. From Corollary 1 and the definition of ξ_λ in (11), we have

$$\begin{aligned} 1 - \Phi(z_\alpha - \xi_{\lambda_0}) &= \lim_{N\to\infty} \beta(c_N|t_{\mathrm{BC}}, N)\\ &= \lim_{N\to\infty} \beta(c_N|t, N')\\ &= \lim_{N\to\infty} \beta\left([N'/N]^{1/2} c_{N'}\big|t, N'\right)\\ &= 1 - \Phi(z_\alpha - \gamma^{1/2}\xi_1). \end{aligned}$$

Thus $\gamma = \xi_{\lambda_0}^2/\xi_1^2$. By (11), if $\lambda_0 \neq 0$, then $(\xi_{\lambda_0}/\xi_1)^2 = (pA_{\lambda_0} + qB_{\lambda_0})^2$. A similar derivation holds for the case when $\lambda_0 = 0$. For the two-sided alternative, the null hypothesis is rejected whenever $t_N^2(\hat{\lambda})$ is large. The proof is the same, but with the normal distribution replaced by a noncentral χ^2 distribution. □

COROLLARY 2. *Suppose the location-shift model* (1) *holds. Then $ARE(t_{\mathrm{BC}}, t) = A_{\lambda_0}^2$.*

COROLLARY 3. *Suppose the scale-shift model* (2) *holds. Then $ARE(t_{\mathrm{BC}}, t) = B_{\lambda_0}^2$.*

3.1. *Bounds on ARE for location shift*

THEOREM 5. *In the case of location shift, suppose conditions* (5)–(10) *hold and $a \leq 1 \leq b$. Then $1 \leq ARE(t_{\mathrm{BC}}, t) \leq \infty$.*

PROOF. By Corollary 2, $\mathrm{ARE}(t_{\mathrm{BC}}, t) = A_{\lambda_0}^2$. Suppose that $\lambda_0 \neq 0$. From the definition of A_{λ_0} in (12), it suffices to show that $\mathrm{Var}(X^{\lambda_0}) \le \lambda_0^2\sigma^2(EX^{\lambda_0-1})^2$. Since $J_0(1) \ge J_0(\lambda_0)$, we have

$$\lambda_0^{-2}\,\mathrm{Var}(X^{\lambda_0})\exp\{-2\lambda_0 E(\log X)\} \le \sigma^2\exp\{-2E(\log X)\}.$$

An application of Jensen's inequality yields

$$\mathrm{Var}(X^{\lambda_0}) \le \lambda_0^2\sigma^2\big[\exp\{(\lambda_0 - 1)E\log X\}\big]^2 \le \lambda_0^2\sigma^2\big(EX^{\lambda_0-1}\big)^2$$

as required. The proof for the case $\lambda_0 = 0$ is similar.

To see that the bounds are sharp, let X be log-norm$(0, \gamma^2)$, that is, $\log(X)$ is $N(0, \gamma^2)$. Then $J_0(\lambda) = \lambda^{-2}\exp(\lambda^2\gamma^2)\{\exp(\lambda^2\gamma^2) - 1\}$ and $\lambda_0 = 0$. Hence $\mathrm{ARE}(t_{\mathrm{BC}}, t) = \gamma^{-2}\exp(2\gamma^2)\{\exp(\gamma^2) - 1\}$ which converges to 1 or ∞ as $\gamma \to 0$ or ∞, respectively. □

3.2. *Bounds on AREs for scale shift.*

THEOREM 6. *In the case of scale shift, suppose that* (5)–(10) *hold and* $a \le 1 \le b$. *Then* $\mathrm{ARE}(t_{\mathrm{BC}}, t) \ge 1$ *if* $\lambda_0 \ge 1$, *and* $0 \le \mathrm{ARE}(t_{\mathrm{BC}}, t) \le \infty$ *otherwise.*

PROOF. Suppose that $\lambda_0 \ge 1$. Since $J_0(\lambda_0) \le J_0(1)$, we have

$$\lambda_0^2\,\mathrm{Var}(X)/\mathrm{Var}(X^{\lambda_0}) \ge \exp\{-2(\lambda_0 - 1)E(\log X)\}.$$

Jensen's inequality gives

$$\mathrm{ARE}(t_{\mathrm{BC}}, t) \ge \{E(X^{\lambda_0})/EX\}^2\{E(X^{\lambda_0-1})\}^{-2}.$$

Since $\mathrm{Cov}(X^{\lambda_0-1}, X) \ge 0$ when $\lambda_0 \ge 1$, we get $E(X^{\lambda_0-1}) \le E(X^{\lambda_0})/EX$, and hence $\mathrm{ARE}(t_{\mathrm{BC}}, t) \ge 1$.

The second part of the theorem is proved by considering the following example. Let $\delta > 0$ and let X have support on the unit interval with density

$$f(x) = \delta x^{\delta-1}, \qquad 0 < x < 1. \tag{13}$$

Then $E(X^\lambda) = \delta/(\delta + \lambda)$, $\mathrm{Var}(X^\lambda) = \delta\lambda^2/(\delta + 2\lambda)(\delta + \lambda)^2$ and $E(\log X) = -\delta^{-1}$. Therefore $\mathrm{ARE}(t_{\mathrm{BC}}, t) = (\delta + 2\lambda_0)/(\delta + 2)$ with $\lambda_0 = \delta/\sqrt{2}$. It follows that $\lim_{\delta\to 0}\mathrm{ARE}(t_{\mathrm{BC}}, t) = 0$.

To see that the upper bound is infinite, let X be log-norm(μ, γ^2). Then $\lambda_0 = 0$, independent of γ^2. In the case of scale shift, $\mathrm{ARE}(t_{\mathrm{BC}}, t) = \{\exp(\gamma^2) - 1\}/\gamma^2 \to \infty$ as $\gamma \to \infty$. □

If one really believes that the data originate from a scale-shift model, it is reasonable to first transform them to a logarithmic scale before applying the t-test, since this transforms a scale-shift model into a location-shift model. Let $t(0)$ denote the t-test applied to the log-transformed data. It is interesting to compare the efficiency of t_{BC} against that of $t(0)$. The next result shows that $\mathrm{ARE}(t_{\mathrm{BC}}, t(0))$ is bounded below by 1 for the scale-shift model.

THEOREM 7. *In the case of scale shift, suppose conditions* (5)–(10) *hold and* $a \le 0 \le b$. *Then* $1 \le ARE(t_{BC}, t(0)) < \infty$.

PROOF. From the proof of Theorem 4, we have $\mathrm{ARE}(t_{BC}, t(0)) = (\xi_{\lambda_0}/\xi_0)^2$, where ξ_λ is defined by (11) with $u = 0$ and $r = 1$. Thus

$$\mathrm{ARE}(t_{BC}, t(0)) = (EX^{\lambda_0})^2 \lambda_0^2 \operatorname{Var}(\log X)/\operatorname{Var}(X^{\lambda_0}).$$

Because λ_0 is the minimizer of $J_0(\lambda)$ on the interval $[a, b]$, which contains 0 by assumption, it follows that $J_0(\lambda_0) \le J_0(0)$, that is,

$$\lambda_0^{-2} \operatorname{Var}(X^{\lambda_0}) \exp\{-2\lambda_0 E \log X\} \le \operatorname{Var}(\log X).$$

But Jensen's inequality implies that $\exp\{E \log X^{\lambda_0}\} \le EX^{\lambda_0}$. Hence

$$\lambda_0^2 \{\operatorname{Var}(X^{\lambda_0})\}^{-1} \operatorname{Var}(\log X)(EX^{\lambda_0})^2$$
$$\ge \lambda_0^2 \{\operatorname{Var}(X^{\lambda_0})\}^{-1} \operatorname{Var}(\log X)(\exp\{E \log X^{\lambda_0}\})^2 \ge 1,$$

proving the theorem. The lower bound of 1 is achieved at the log-normal distribution. It can be shown that for the gamma distribution with density function

$$(14) \qquad f(x) = \beta^\theta x^{\theta - 1} \exp(-\beta x)/\Gamma(\theta), \qquad x > 0, \theta, \beta > 0,$$

that $\lambda_0 \to 1/3$ and $\mathrm{ARE}(t_{BC}, t(0)) \to 1$ as $\theta \to \infty$, independent of β. □

4. Numerical results. Numerical values of the AREs are computed for the following distributions.

Gamma distribution. Tables 1 and 2 give values for the gamma density (14). The AREs are scale invariant and so are independent of β. The AREs of t_{BC} to the Wilcoxon and normal scores rank tests are included in Table 1 for comparison. The reason that the ARE of the Box–Cox t-test to the normal scores test is 0 when $\theta = 1$ is because the efficacy of the latter test is infinite at the exponential distribution.

The low efficiency of t_{BC} versus $t(0)$ may be explained as follows. If the two populations differ by a location shift, then the support of the distribution of $\log(X)$ is the whole real line, whereas that of $\log(Y)$ is a half-line. If the left

TABLE 1
AREs for gamma distributions under location shift

θ	λ_0	$\mathrm{ARE}(t_{BC}, t)$	$\mathrm{ARE}(t_{BC}, t(0))$	$\mathrm{ARE}(t_{BC}, W)$	$\mathrm{ARE}(t_{BC}, NS)$
1.0	0.265	11.44	0	3.81	0
1.5	0.289	3.12	0.486	1.71	0.898
2.0	0.301	2.12	0.683	1.41	0.967
5.0	0.321	1.28	0.907	1.14	0.998
∞	1/3	1	1	$\pi/3$	1

Note: W and NS denote the Wilcoxon and normal scores tests, respectively.

TABLE 2
AREs for gamma distributions under scale and location–scale shift ($r = u = 1$)

θ	λ_0	Scale shift ARE(t_{BC}, t)	Scale shift ARE($t_{BC}, t(0)$)	Location–scale ARE(t_{BC}, t)
1.0	0.265	0.803	1.32	4.58
1.5	0.289	0.864	1.21	1.60
2.0	0.301	0.896	1.16	1.25
5.0	0.321	0.957	1.06	1.01
∞	1/3	1	1	1

tail of the distribution of $\log(X)$ is sufficiently thick, as is apparently the case here, it would be easier for the $t(0)$-test to detect a difference in means than it is for the t_{BC}-test.

Log-normal distribution. The AREs when X is log-normal(μ, γ^2) are given in Tables 3 and 4. These AREs are independent of μ. The value of ARE(t_{BC}, t) increases with the skewness parameter γ^2. Since $\lambda_0 = 0$, we have ARE$(t_{BC}, t(0)) = 1$ here.

Uniform distribution on $(0, h)$. Assume without loss of generality that $h = 1$. Then $J_0(\lambda) = \exp(2\lambda)/\{(2\lambda + 1)(\lambda + 1)^2\}$, $\lambda_0 = 1/\sqrt{2}$, and for location shift, we have ARE$(t_{BC}, t) = 1.173$ and ARE$(t_{BC}, t(0)) = 0$. For scale shift, the corresponding values are ARE$(t_{BC}, t) = 0.802$ and ARE$(t_{BC}, t(0)) = 2.41$.

TABLE 3
AREs for log-normal distributions under location shift

	γ^2 = 0.10	0.50	0.75	1.00
ARE(t_{BC}, t)	1.29	3.53	6.68	12.70
ARE(t_{BC}, W)	1.10	1.35	1.52	1.73

Note: W denotes the Wilcoxon test; ARE$(t_{BC}, t(0)) = 1$ in this case because $\lambda_0 = 0$.

TABLE 4
ARE(t_{BC}, t) for log-normal distributions under scale and location–scale shift ($r = u = 1$)

	γ^2 = 0.10	0.50	0.75	1.00
Scale	1.05	1.30	1.49	1.72
Location–scale	1.16	2.14	3.15	4.68

Note: ARE$(t_{BC}, t(0)) = 1$ in this case because $\lambda_0 = 0$.

TABLE 5
AREs for inverse Gaussian (Wald) distributions under location shift

ϕ	λ_0	ARE(t_{BC}, t)	ARE(t_{BC}, W)	ARE(t_{BC}, NS)	ARE(t_{BC}, $t(0)$)
2.0	−0.0502	2.87	1.31	0.963	1.04
3.0	−0.0383	2.10	1.23	0.980	1.02
4.0	−0.0304	1.77	1.19	0.987	1.01
5.0	−0.0256	1.59	1.17	0.991	1.01

Note: ϕ denotes the shape parameter; W and NS are the Wilcoxon and normal scores tests, respectively.

Inverse Gaussian (Wald) distribution. Table 5 gives AREs under location shift for the inverse Gaussian distribution with shape parameter ϕ and density function

$$f(x) = (\phi/2\pi)^{1/2} x^{-3/2} \exp\{-(1/2)\phi(x-1)^2/x\}, \qquad x > 0, \phi > 0.$$

The fact that all the values of ARE(t_{BC}, $t(0)$) exceed 1 in the table is not a coincidence, as the following theorem shows.

THEOREM 8. *In the case of location shift, suppose conditions* (5)–(10) *hold and that* $a \le 0 \le b$ *and* $\lambda_0 < 0$. *Then* $ARE(t_{BC}, t(0)) \ge 1$.

PROOF. The definition of λ_0 implies that $J_0(0) \ge J_0(\lambda_0)$, which gives

$$\mathrm{Var}(\log X) \ge \lambda_0^{-2}\,\mathrm{Var}(X^{\lambda_0}) \exp(-2\lambda_0 E \log X).$$

From the assumption that $\lambda_0 < 0$, we have $\mathrm{Cov}(X^{\lambda_0}, X^{-1}) \ge 0$, which by Jensen's inequality yields

$$EX^{\lambda_0 - 1} \ge EX^{\lambda_0} EX^{-1} \ge \exp(\lambda_0 E \log X) EX^{-1}.$$

Therefore by Corollary 2,

$$\begin{aligned} \mathrm{ARE}(t_{BC}, t(0)) &= (A_{\lambda_0}/A_0)^2 \\ &= \lambda_0^2\,\mathrm{Var}(\log X)\big(EX^{\lambda_0-1}/EX^{-1}\big)^2/\mathrm{Var}(X^{\lambda_0}) \ge 1. \qquad \square \end{aligned}$$

In view of the preceding result and the small values of ARE(t_{BC}, $t(0)$) in Table 1, it is natural to ask whether the ARE can be greater than 1 for some distribution with $\lambda_0 > 0$. Such an example is provided by the distribution with density function (13). Assume that δ is large so that the moment conditions are satisfied. Direct computation gives $\lambda_0 = \delta/\sqrt{2}$ and

$$\begin{aligned} \mathrm{ARE}(t_{BC}, t(0)) &= (1 + 2/\sqrt{2})(1 + 1/\sqrt{2})^2\big[(\delta - 1)/\big(\delta(1 + 1/\sqrt{2}) - 1\big)\big]^2 \\ &\to 1 + \sqrt{2}, \qquad \delta \to \infty. \end{aligned}$$

5. Simulation results. A simulation experiment was carried out to compare the small-sample performance of the tests for the case of log-normal distributions under location shift. Table 6 gives the simulated levels and powers of the t_{BC}, t and W tests for $(m, n) = (20, 30)$ and $(30, 50)$, at nominal

TABLE 6

Simulated levels and powers of the Box–Cox t-test (t_{BC}), Wilcoxon rank test (W) and the ordinary t-test (t)

			γ^2			
			0.10	**0.50**	**0.75**	**1.00**
Shift = 0.0	$m = 20$	t_{BC}	0.056	0.058	0.056	0.058
	$n = 30$	$(E\hat{\lambda})$	(−0.008)	(−0.003)	(−0.004)	(−0.003)
		t	0.054	0.051	0.048	0.046
		W	0.051	0.051	0.050	0.049
	$m = 30$	t_{BC}	0.055	0.053	0.055	0.053
	$n = 50$	$(E\hat{\lambda})$	(−0.009)	(−0.005)	(−0.004)	(−0.004)
		t	0.054	0.049	0.049	0.045
		W	0.050	0.048	0.049	0.049
Shift = 0.1	$m = 20$	t_{BC}	0.224	0.106	0.097	0.093
	$n = 30$	$(E\hat{\lambda})$	(−0.027)	(−0.050)	(−0.053)	(−0.056)
		t	0.186	0.069	0.055	0.047
		W	0.122	0.068	0.064	0.059
	$m = 30$	t_{BC}	0.318	0.131	0.119	0.115
	$n = 50$	$(E\hat{\lambda})$	(−0.029)	(−0.055)	(−0.057)	(−0.059)
		t	0.259	0.071	0.058	0.050
		W	0.167	0.079	0.070	0.070
Shift = 0.2	$m = 20$	t_{BC}	0.627	0.240	0.204	0.186
	$n = 30$	$(E\hat{\lambda})$	(0.029)	(−0.057)	(−0.064)	(−0.068)
		t	0.543	0.121	0.081	0.063
		W	0.345	0.121	0.101	0.088
	$m = 30$	t_{BC}	0.812	0.340	0.286	0.269
	$n = 50$	$(E\hat{\lambda})$	(0.028)	(−0.062)	(−0.068)	(−0.070)
		t	0.725	0.148	0.096	0.070
		W	0.497	0.165	0.136	0.123
Shift = 0.3	$m = 20$	t_{BC}	0.913	0.432	0.354	0.317
	$n = 30$	$(E\hat{\lambda})$	(0.130)	(−0.038)	(−0.052)	(−0.059)
		t	0.858	0.211	0.131	0.092
		W	0.631	0.210	0.166	0.140
	$m = 30$	t_{BC}	0.985	0.600	0.505	0.461
	$n = 30$	$(E\hat{\lambda})$	(0.129)	(−0.042)	(−0.055)	(−0.061)
		t	0.962	0.279	0.161	0.110
		W	0.821	0.306	0.242	0.209

Note: The populations were log-normal$(0, \gamma^2)$ under location shift. The nominal level for all tests was 0.05, and 40,000 Monte Carlo trials were performed, giving a maximum simulation standard error of 0.0025. For each sample size, the same simulated samples were used to obtain the level and power of the tests. Average values of $\hat{\lambda}$ are given in parentheses.

TABLE 7
Simulated powers of the Box–Cox t-test (t_{BC}), Wilcoxon rank test (W) and the ordinary student t-test (t) for sample sizes adjusted according to Table 3

	γ^2			
	0.10	**0.50**	**0.75**	**1.00**
t_{BC}	0.63	0.24	0.21	0.19
(m,n)	(20, 30)	(20, 30)	(20, 30)	(20, 30)
t	0.63	0.26	0.23	0.22
(m,n)	(26, 38)	(71, 106)	(134, 200)	(254, 381)
W	0.37	0.14	0.13	0.12
(m,n)	(22, 33)	(27, 41)	(31, 46)	(35, 52)

Note: The populations used are log-normal(0, γ^2) with location shift 0.2. The nominal level for all three tests is 0.05. Each result is based on 10,000 Monte Carlo trials, giving a maximum simulation standard error of 0.005.

level 0.05. Estimated values of $E\hat{\lambda}$ are given in parentheses. The $t(0)$-test is not included because it is asymptotically equivalent to t_{BC} since $\lambda_0 = 0$. The following conclusions may be summarized from the table:

1. The significance levels of the t_{BC}-test seem to be somewhat high when $(m, n) = (20, 30)$, although they get closer to the nominal values (within two simulation standard errors) when (m, n) is increased to $(30, 50)$.
2. The power of the t_{BC}-test dominates that of the other two tests uniformly over the cases considered, by as much as more than four times the power of the t-test when $(m, n) = (30, 50)$, $\gamma^2 = 1$, and the location shift is 0.3 units.

To check on the relevance of the AREs in Table 3 for finite-sample sizes, another simulation was conducted with $m = 20$ and $n = 30$ for t_{BC} in the log-normal location-shift case. The sample sizes of the t- and W-tests were modified according to the AREs in that table. Specifically, the sample sizes for the t-test were $(m, n) = (20\,\mathrm{ARE}(t_{BC}, t), 30\,\mathrm{ARE}(t_{BC}, t))$, and those for the W-test were $(m, n) = (20\,\mathrm{ARE}(t_{BC}, W), 30\,\mathrm{ARE}(t_{BC}, W))$. If the asymptotic formulas hold at these sample sizes, the powers of the three tests should be about the same. Table 7 gives the results which can be summarized as follows:

1. The formula for $\mathrm{ARE}(t_{BC}, t)$ is quite accurate for the values of γ^2 shown in the table.
2. The formula for $\mathrm{ARE}(t_{BC}, W)$ underpredicts the relative efficiency at the sample sizes considered, that is, the t_{BC}-test is more efficient relative to the Wilcoxon test than their ARE suggests.

6. More than two samples. The efficiency results for the two-sample Box–Cox t-test are extended to the one-way analysis of variance F-test in this section. For each i, let $X_{i1}, \dots, X_{in_i}$ be independent copies of a positive variable X, which satisfies conditions analogous to (5)–(10). Given $c_i = \tau_i n_i^{-1/2} \geq 0$, $u_i \geq 0$ and $r_i \geq 0$ with $(\tau_i, u_i, r_i) \neq (\tau_j, u_j, r_j)$ for some (i, j), define $Y_{ij} = (1 + c_i)^{r_i}(X_{ij} + c_i u_i)$, $j = 1, \dots, n_i$, $i = 1, \dots, I$. Let $N = \sum_{i=1}^I n_i$ and assume that there exist $k_1, \dots, k_I$ such that $\sum_{i=1}^I k_i = 1$ and $n_i/N \to k_i \in (0, \infty)$, $i = 1, \dots, I$. Consider testing the null hypothesis H_0: $\tau_1 = \cdots = \tau_I = 0$ versus local alternatives with $\tau_i > 0$ for some i. Let $\bar{Y}_i(\lambda) = n_i^{-1}\sum_{j=1}^{n_i} Y_{ij}(\lambda)$, $\tilde{Y}(\lambda) = N^{-1}\sum_{i=1}^I n_i \bar{Y}_i(\lambda)$, $\text{RSS}_H(\lambda) = \sum_{i=1}^I n_i\{\bar{Y}_i(\lambda) - \tilde{Y}(\lambda)\}^2$ and

$$S_N^2(\lambda) = \frac{1}{N - I}\sum_{i=1}^{I}\sum_{j=1}^{n_i}\left\{Y_{ij}(\lambda) - \bar{Y}_i(\lambda)\right\}^2.$$

The ANOVA F-statistic based on the Box–Cox transformed data is given by

$$F_N(\lambda) = (I - 1)^{-1}\text{RSS}_H(\lambda)/S_N^2(\lambda).$$

Let $\log(\dot{z}) = N^{-1}\sum_{i=1}^I\sum_{j=1}^{n_i}\log(Y_{ij})$, $\sigma^2(\lambda) = \text{Var}\{X(\lambda)\}$, $\eta = E\log(X)$ and define $J_N(\lambda) = S_N^2(\lambda)/\dot{z}^{2\lambda}$ and $J_0(\lambda) = \sigma^2(\lambda)/e^{2\lambda\eta}$. Finally, let $\hat{\lambda}$ and λ_0 be the minimizers of $J_N(\lambda)$ and $J_0(\lambda)$, respectively. The Box–Cox F-test rejects H_0 if $F_N(\hat{\lambda}) \geq F_{I-1, N-I; \alpha}$, where $F_{I-1, N-I; \alpha}$ is the upper α-quantile of the F-distribution with $I - 1$ and $N - I$ degrees of freedom.

We collect together here generalizations of the previous results to multisamples. The proofs are omitted since they are similar to those for the case of two samples.

THEOREM 9. *$\hat{\lambda} \to \lambda_0$ a.s. and $N^{1/2}(\hat{\lambda} - \lambda_0) = O_P(1)$.*

THEOREM 10. *$F_N(\hat{\lambda}) - F_N(\lambda_0) \to 0$ in probability.*

THEOREM 11. *Let $\xi_i = \tau_i(u_i EX^{\lambda_0 - 1} + r_i EX^{\lambda_0})$, $\bar{\xi} = \sum_{i=1}^I k_i \xi_i$ and*

$$\delta^2(\lambda_0) = \left[\text{Var}\{X(\lambda_0)\}\right]^{-1}\sum_{i=1}^{I} k_i\left(\xi_i - \bar{\xi}\right)^2. \tag{15}$$

Then $F_N(\hat{\lambda}) \to_D (I - 1)^{-1}\chi^2_{I-1}(\delta^2(\lambda_0))$ under the null as well as alternative hypotheses.

THEOREM 12. *The asymptotic relative efficiency of the Box–Cox F-test, F_{BC}, against the ordinary F-test, F, is $ARE(F_{\text{BC}}, F) = \delta^2(\lambda_0)/\delta^2(1)$, where $\delta^2(\lambda_0)$ is defined in* (15).

COROLLARY 4 (Location shift). *Assume that $r_1 = \cdots = r_I = 0$. Then*

$$ARE(F_{\text{BC}}, F) = \text{Var}(X)(EX^{\lambda_0 - 1})^2/\text{Var}\{X(\lambda_0)\}.$$

COROLLARY 5 (Scale shift). *Assume that $u_1 = \cdots = u_I = 0$. Then*

$$ARE(F_{\text{BC}}, F) = \text{Var}(X)(EX^{\lambda_0}/EX)^2/\text{Var}\{X(\lambda_0)\}.$$

Since the AREs are the same as those in the two-sample case for both location and scale shifts, the previous bounds apply. For example, consider the following result.

THEOREM 13. *In the case of location shift or scale shift (the latter with* $\lambda_0 \geq 1$), $1 \leq ARE(F_{BC}, F) < \infty$.

7. Discussion. The purpose of this paper is to provide some answers to the practical question of whether one should first transform the data before carrying out a two-sample t-test. Based on the theoretical and simulation results reported here, the answer is clearly "yes," at least for location-shift models.

It may seem strange that we are advocating a transformation of the data before applying the t-test in the case of location shift, because on the transformed scale the model is no longer one of location shift, whereas the t-test is designed for the latter. Two reasons may explain why our results are counter to this kind of intuition.

1. While the t-test is designed for location-shift alternatives, it is really efficient only when the populations are normal. The literature on rank tests, for example, contains numerous examples of alternative tests that are superior to the t-test, sometimes by a large margin, for nonnormal populations. In the situations considered in this paper, where the data are distributed on the positive real line, the extent of deviation from normality may be large.
2. Because power transformations are monotone, the transformed t-test is consistent even if the model after transformation is not one of location shift. If we recall that the Box–Cox transformation is chosen to make the transformed data satisfy the normality and homoscedasticity assumptions as closely as possible, it should not be surprising that the t_{BC}-test performs satisfactorily against the t- or $t(0)$-tests under the location-shift and scale-shift models, respectively, for nonnormal populations.

REFERENCES

ATKINSON, A. C. (1985). *Plots, Transformations and Regression*. Oxford Univ. Press.

BICKEL, P. J. and DOKSUM, K. A. (1981). An analysis of transformations revisited. *J. Amer. Statist. Assoc.* **76** 296–311.

BOX, G. E. P. and COX, D. R. (1964). An analysis of transformations (with discussion). *J. Roy. Statist. Soc. Ser. B* **26** 211–246.

BOX, G. E. P., HUNTER, W. G. and HUNTER, J. S. (1978). *Statistics for Experimenters*. Wiley, New York.

CARROLL, R. J. (1980). A robust method for testing transformations to achieve approximate normality. *J. Roy. Statist. Soc. Ser. B* **42** 71–78.

DOKSUM, K. A. (1987). An extension of partial likelihood methods for proportional hazard models to general transformation models. *Ann. Statist.* **15** 325–345.

DOKSUM, K. A. and WONG, C.-W. (1983). Statistical tests based on transformed data. *J. Amer. Statist. Assoc.* **78** 411–417.

HERNANDEZ, F. and JOHNSON, R. A. (1980). The large-sample behavior of transformations to normality. *J. Amer. Statist. Assoc.* **75** 855–861.

HINKLEY, D. V. (1975). On power transformations to symmetry. *Biometrika* **62** 101–111.

HINKLEY, D. V. and RUNGER, G. (1984). The analysis of transformed data (with discussion). *J. Amer. Statist. Assoc.* **79** 302–309.

RUBIN, H. (1956). Uniform convergence of random functions with applications to statistics. *Ann. Math. Statist.* **27** 200–203.

DEPARTMENT OF MATHEMATICS AND STATISTICS
BOWLING GREEN STATE UNIVERSITY
BOWLING GREEN, OHIO 43403

DEPARTMENT OF STATISTICS
UNIVERSITY OF WISCONSIN
MADISON, WISCONSIN 53706