Theor Appl Genet (1993) 87:487–497

THEORETICAL AND APPLIED GENETICS

Needs and benefits of empirical power transformations for production and quality traits in forest tree breeding

G. Jansson, Ö. Danell

The Forestry Research Institute of Sweden, Glunten, S-75183 Uppsala, Sweden

Received: 23 April 1993 / Accepted: 3 May 1993

Abstract. Non-normality in the distribution of individual observations of production and quality traits in forest tree breeding may cause inaccurate selection and overestimation of predicted selection gain. The distribution of individual observations of traits such as height, diameter, branch diameter, branch angle and number of branches per whorl is not always normal. We investigated how the observations were distributed and to what degree it is possible to improve normality, homogeneity of error variance and additivity by using empirical power transformations. Computer simulations showed that a seriously skewed distribution impairs selection efficiency and exaggerates selection gain expectations. If the distribution is heavily skewed, transformation might be worthwhile. It does not seem possible to offer any general advice about which variates should be transformed, but in most cases there seems to be no need of any transformation.

Key words: Forest tree breeding – Non-normality – Box-Cox transformations – Genetic evaluation – Selection efficiency

Introduction

In progeny trials, variates such as height, diameter, branch diameter, branch angle and number of branches per whorl are often chosen for assessment mainly from aspects concerning ease of assessment. In the evaluation of progeny trials, these production and quality variates have sometimes shown non-normal distributions. The effects of non-normality in the distribution of individual observations of production and quality traits on selection and estimation of genetic gain have attracted very little attention in the area of forest tree breeding.

In genetic evaluations, linear models are used as they have certain advantages regarding such matters as calculations and interpretation. When using linear models, we usually assume (1) normality of distribution, (2) homogeneity of error variance, (3) additivity and (4) independence of observations. The breaking of these conditions can lead to inaccurate selection and a loss of efficiency in the estimation of effects. The assumption of homogeneity of variance is also difficult to fulfil because fast-growing families are often more variable than slow-growing families. The distributions of individual observations often have a heavy upper tail. However, even if the individual observations are non-normally distributed, the mean values may be approximately normal, according to 'the central limit theorem'. Thus, plot means might be sufficiently normal, unless the observations per plot are few.

If assumptions 1–3 are not satisfied, a transformation of the original observations may improve the situation. Transformations such as logarithmic, angular, square root, etc. have been examined, and investigators have suggested that they achieve normality and homogeneity of variances (e.g. Bartlett 1947). Such studies have concentrated on obtaining constant error variances. Box and Cox (1964) suggested a procedure for estimating the best transformation within the family of power transformations. Their aim was to get the observations normally distributed with constant variances. The transformation should be chosen partly on the basis of information provided by the data and

Communicated by P. M. A. Tigerstedt
Correspondence to: G. Jansson

partly on general considerations of simplicity and ease of interpretation. A retransformation is needed to refer back to the original scale, but this will not as a rule give unbiased estimators (e.g. Neyman and Scott 1960).

Ibe and Hill (1988) used power transformations of poultry egg production data to improve normality, homoscedasticity and linearity of the genotypic regressions. They found that a power transformation of egg production data produced both normality of distribution and homogeneity of variance. The transformation also improved the linearity of genotypic regressions and usually gave higher heritabilities. Their conclusion was that the transformation predicts the genetic worth of animals more efficiently, so that more efficient selection decisions can be made.

Kung (1988) discussed the need for transformation is forest genetics. His conclusion was that transformations are of advantage for simplifying relationships, for stabilizing variance and improving normality, but that the transformed model should also be understandable and practical.

Most traits of commercial interest in forest tree breeding have an assumed polygenic background, which makes the expected distribution of individual observations approximately normal. Therefore, a normally distributed 'underlying' variate value is justified on theoretical grounds. For instance, skewness may be considered to be introduced into the physiological expressions of the underlying genetic values. Besides the mathematical advantages of using linear models, there are biological reasons to strive for normality in variate values used in genetic evaluation and selection, but there is no theoretical guidance on how to achieve this. The alternatives are then to find the best transformation from the data itself or to use a standard transformation such as logarithmic, square root, etc.

The first objective of the study presented here was to investigate the distributed properties of individual observations for production and quality variates. For variates with non-normal distributions, a second objective was to estimate empirical power transformations that improve the distribution of the data to meet the first three assumptions mentioned above (normal distribution, homogeneity of error variance and additivity). A third objective was to study how selection, efficiency and expectancy of genetic gain are affected if the variates are not normally distributed. Observed trial data were used for the first and second objectives and simulated data for the third objective.

Material and methods

Data were obtained from ten progeny trials with Scots pine (*Pinus sylvestris* L.) in central Sweden. The trials were analysed as unbalanced randomized blocks with single-tree plots. The trials had attained an age between 5 and 18 years at measurement and mean heights between 118 and 586 cm. Table 1 lists the variates measured. Volume was estimated with volume functions, using height and diameter as independent variables. Different functions were used for trees with a breast height diameter below (Andersson 1954) and above 5 cm (Näslund 1947).

Table 1. Description of measured variates in the trial datasets and the corresponding abbreviations

Measured variate	Unit	Abbreviation
Total height	cm	H1
Height 3–7 years earlier	cm	H2
Height increment, H1-H2	cm	INC
Breast height diameter	mm	DIA
Volume	dm^3	VOL
Branch diameter, thickest branch (whorl 1.5 m)	mm	BD1
Branch diameter, opposite to the thickest branch	mm	BD2
Branch angle, thickest branch	Angle degrees	BA1
Branch angle, opposite to the thickest branch	Angle degrees	BA2
Number of branches		BNUM

Finding transformations

The following fixed model was applied to find a suitable transformation:

$$y_{ij} = \mu + b_i + \varepsilon_{ij} \quad (1)$$

where

y_{ij} = observed value for tree ij
μ = grand mean
b_i = fixed effect of block i
ε_{ij} = residual effect for tree ij, $\sim (0, \sigma_\varepsilon^2)$

A more appropriate model would have been $y_{ijk} = \mu + b_i + f_{ij} + e_{ijk}$, where f_{ij} is the random effect of family j in block i. However, without replications of families within blocks this model cannot be used if we want to study variance, skewness and kurtosis for the residuals within each family. In model (1) the residual is a composite of the two random effects, family and a 'pure' error. Thus, one can consider the residuals as a random variable, but the residuals ε are not completely independent of each other. The REML method in the SAS 'Varcomp' procedure was used to estimate variance components (SAS 1990).

To estimate the best transformation within the family of power transformations, an empirical maximum likelihood procedure developed by Box and Cox (1964) was applied in this study. The first step was to obtain the transformation

$$y^{(\lambda)} = \begin{cases} \dfrac{y^\lambda - 1}{\lambda} & (\lambda \neq 0) \\ \ln y & (\lambda = 0) \end{cases} \quad (2)$$

Since an analysis of variance is not affected by a linear transformation, this is equivalent to

$$y^{(\lambda)} = \begin{cases} y^\lambda & (\lambda \neq 0) \\ \ln y & (\lambda = 0) \end{cases} \quad (3)$$

The first form is preferable for theoretical analysis because it is continuous at $\lambda = 0$ and has, therefore, been used here.

Transformations were made for different values for λ, and model (1) was then applied to estimate the residuals for the different λ. Setting $\lambda = 1$ gives no change, $\lambda = 0.5$ is equivalent to a square root transformation, $\lambda = 0$ means a logarithmic transformation and $\lambda = -1$ is equivalent to a reciprocal transformation. Other values for λ give transformations between these standard alternatives.

The second step was to find the value of λ that maximizes the log likelihood function

$$L_{\max}(\lambda) = -\frac{v \ln s_\lambda^2}{2} + \frac{(\lambda - 1)v}{n} \sum \ln y \tag{4}$$

where

$L_{\max}(\lambda)$ = log likelihood function
n = number of observations
s_λ^2 = residual variance of transformed y values from model (1)
v = degrees of freedom for s_λ^2
$\sum \ln y$ = sum of logarithms of the observed values.

To find the λ that maximizes L, a second-degree polynomial was fitted to describe the relationship between λ and L:

$$L = k_1 + k_2 \lambda + k_3 \lambda^2 \tag{5}$$

where k_1, k_2 and k_3 are regression parameters.

The maximum value of L was then found by setting the derivative of this function to zero and solving for λ, which gives:

$$\lambda = \frac{k_2}{2k_3} \tag{6}$$

Confidence intervals were computed from the Box-Cox algorithm in order to test whether the transformation parameter λ differed from $\lambda = 1$. The value $-2L_{\max}$ follows approximately the χ^2-distribution. The upper and lower confidence limits, λ_1 and λ_2, correspond to the pair of values of λ that have L values equal to

$$L = L_{\max} - \tfrac{1}{2}\chi^2_{\alpha(1)} \tag{7}$$

Solutions for λ_1 and λ_2 were obtained by finding the roots of (5).

A method suggested by Sokal and Rohlf (1981) with which to achieve normality and induce homogeneity of variances at the same time was also tried. The method is a composite of (4) and Bartlett's test of homogeneity of variances. A new log-likelihood function is defined as

$$L' = L - \tfrac{1}{2}\chi^2 \tag{8}$$

where L is as defined in (4) and χ^2 is Bartlett's test of homogeneity of variances computed as

$$\chi^2 = \left\{ \left[\sum_{i=1}^{r} (n_i - 1) \right] \ln s^2 - \sum_{i=1}^{r} (n_i - 1) \ln s_i^2 \right\} \Big/ C \tag{9}$$

where

$$s^2 = \frac{\sum_{i=1}^{r} (n_i - 1) s_i^2}{N - r}$$

$$C = 1 + \frac{1}{3(r-1)} \left[\sum_{i=1}^{r} \frac{1}{(n_i - 1)} - \frac{1}{\sum_{i=1}^{r} (n_i - 1)} \right]$$

and

n_i = number of observations in sample i
s_i^2 = sample variance
N = total number of observations
r = number of samples
C = correction factor.

Checking assumptions on the transformed scale

After the best-fitted transformation had been applied to the data, normality and homogeneity of variances were checked. Deviation from normality was measured by skewness and kurtosis. First, in order to standardize the values and make untransformed and transformed values comparable, we made the following calculations on the residuals from model (1) and on both untransformed and transformed values:

$$x = \frac{p_t - m_t}{s_t} \tag{10}$$

$$y_u = \frac{r_u}{s_u} \tag{11}$$

$$y_t = \frac{r_t}{s_t} \tag{12}$$

where

x = transformed value for each block expressed as deviation from the mean of the blocks
p_t = predicted value of the transformed variable
m_t = mean of the transformed variable
s_t = square root of the mean square error of the transformed value
y_u = residual of the untransformed observation expressed in standard deviations
r_u = residual of the untransformed observation
s_u = square root of the mean square error of the untransformed value
y_t = residual of the transformed observation expressed in standard deviations
r_t = residual of the transformed observation.

Second, linear regressions for variance, skewness and kurtosis of the x-values, as calculated above, on y_u or y_t were calculated to cheek the effects of transformation. Means and the regression coefficients were calculated both for the original values and the transformed. A regression coefficient of zero denotes absence of any trend between high and low values.

Bartlett's test for homogeneity of variances was used to measure the effect on homogeneity of variances in the transformations.

Simulation study

Computer simulations were used to measure the effects of different distributions on evaluation procedure and selection efficiency in terms of achieved genetic gain. Simulations were made for variables distributed as $\sqrt{y}, y, y^2, y^5, y^{10}$ and $\exp(y)$, i.e. these are variate values for which Box-Cox transformations with $\lambda = 2, 1, 0.5, 0.2, 0.1$ and 0 are the appropriate ones. The variable y represents a normally distributed variate. As the exponential distribution is dependent on the level of the values, these transformations should be seen just as special examples. For the power transformation it is the relation rather than the level of the values that affects the result, which means that these results are general.

The simulations were exemplified for two situations, one with a half-sib family structure (progeny testing) and one with clonal propagation of the individual genotypes (clonal test). Two levels of genetic variation were chosen to cover the most common interval of the genetic coefficient of variation found in Swedish test data with different variates. Twenty-five data sets for these two situations were generated according to the following models:

$$\text{Half-sib data: } y_{ijk} = \mu + b_i + s_j\sqrt{\tfrac{1}{4}\sigma_A^2} + r_{ijk}\sqrt{\sigma_p^2 - \tfrac{1}{4}\sigma_A^2} \quad (13)$$

$$\text{Clone data: } y_{ijk} = \mu + b_i + c_j\sqrt{\sigma_G^2} + r_{ijk}\sqrt{\sigma_P^2 - \sigma_G^2} \quad (14)$$

where

y_{ijk} = value for observation ijk
μ = grand mean, set to 8
b_i = fixed effect of block i, $i = 1,\ldots,8, \sum b_i = 0, b_{\min} = -1.25$, $b_{\max} = 1.50$
s_j = a random number for effect of parent j, $\sim NID(0,1)$
c_j = a random number for effect of clone j, $\sim NID(0,1)$
r_{ijk} = a random number for each observation ijk, $\sim NID(0,1)$
σ_A^2 = additive genetic variance ($= h^2$ when $\sigma_P^2 = 1$)
σ_G^2 = genotypic variance ($= H^2$ when $\sigma_P^2 = 1$)
σ_P^2 = phenotypic variance (set to 1)
h^2 = narrow sense heritability, i.e. σ_A^2/σ_P^2
H^2 = broad sense heritability or 'repeteability', i.e. σ_G^2/σ_P^2.

The assumptions used in the simulations are shown in Table 2.

The unbalancedness in the data was created by giving all trees a random number between 0 and 1, from a uniform distribution, and then deleting all trees with a value above 0.8, which corresponds to an expected survival rate of 80%.

Variance components were estimated in each of the generated datasets using the REML method in the SAS 'Varcomp' procedure (SAS 1990). Means of the estimated variance components in the 25 datasets were used as prior variances when calculating BLUP of family (parent) and clonal genetic values, respectively, from Henderson's Mixed Model Equations (Henderson 1975) in each of the 25 datasets. The models are in matrix notation

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Zu} + \mathbf{e} \quad (15)$$

where

$\mathbf{y}$ = a vector with untransformed and transformed observed values, respectively
$\mathbf{X}$ = a design matrix of fixed effects
$\mathbf{Z}$ = a design matrix of random effects
$\mathbf{b}$ = a vector of fixed block effects
$\mathbf{u}$ = a vector of random family or clonal effects, respectively
$\mathbf{e}$ = a vector with residual effects.

$\hat{\mathbf{b}}$ and $\hat{\mathbf{u}}$ are solved as

$$\begin{bmatrix}\hat{\mathbf{b}}\\ \hat{\mathbf{u}}\end{bmatrix} = \begin{bmatrix}\mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z}\\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z} + \mathbf{I}\frac{\sigma_e^2}{\sigma_u^2}\end{bmatrix}^{-1}\begin{bmatrix}\mathbf{X}'\mathbf{y}\\ \mathbf{Z}'\mathbf{y}\end{bmatrix} \quad (16)$$

where

σ_e^2 = residual variance
σ_u^2 = family or clone variance, respectively.

Parents and clones, respectively, were then ranked and selected in various proportions. The efficiency of the selection and the effect on the genetic gain from the transformation were calculated in the following way.

Table 2. Assumptions used in the simulations of test data

	Half-sib families	Clones
Number of half-sib families or clones	40	40
Number of blocks	8	8
Number of seedlings per genetic entry	40	8
Original number of seedling per block or ramets per clone and block	5	1
Survival rate (average)	80%	80%
Genetic coefficient of variation[a]	5.6% and 11.2%	6.2% and 12.5%
Heritability	$h^2 = 0.20$	$H^2 = 0.25$
Number of datasets	25	25

[a] $\frac{\sigma_A}{\mu}$ for half-sib data, $\frac{\sigma_G}{\mu}$ for clone data, where σ_A and σ_G are additive genetic and genotypic standard deviations, respectively

$$\text{Actual selection efficiency} = \frac{\hat{I}_n(\text{rank } s)}{\hat{I}_n(\text{rank } n)} \quad (17)$$

$$\text{Ratio of the predicted gains} = \frac{\hat{I}_s(\text{rank } s)}{\hat{I}_s(\text{rank } n)} \quad (18)$$

where

$\hat{I}_n$ = mean BLUP of parent or clone effect of the selected part in the normally distributed data
$\hat{I}_s$ = mean BLUP of parent or clone effect of the selected part within the generated skewed distribution of y
rank n indicates ranking obtained on BLUP computed from normally distributed data
rank s indicates ranking obtained on BLUP computed with the skewed distribution

The ratio of the predicted gains measures the bias in gain expectations.

Results

Transformations required for the empirical data

Table 3 shows the transformation parameters (λ) that maximized the log-likelihood for different variates in the different trials. The disparity between results with the Box-Cox transformation and the modified procedure given by Sokal and Rohlf (1981) was small, which is why only the results from the Box-Cox transformation are presented. The Box-Cox procedure appeared simpler to use. From Table 3 it is not possible to discern any general pattern in the transformation parameters. Transformation parameters were plotted versus the mean height, but they too displayed no specific pattern of relationships for any of the variates. Generally speaking, height, height increment and diameter seem to be variates with no need for transformation, whereas volume, branch diameter, branch angle and number of branches sometimes seem to be worth a

Table 3. Transformation parameters which maximize the log-likelihood function for different variates and trails. Variate abbreviations are given in Table 1

Trial code	Mean height (cm)	Measured variates									
		H1	H2	INC	DIA	VOL	BD1	BD2	BA1	BA2	BNUM
101A	470	1.02	0.44	1.13	0.74	0.27	0.12			0.40	0.24
101C	392	0.11			0.27	0.03	0.28		−0.11		0.16
101D	586	1.68			1.21	0.47					
101E	484	1.17			1.01	0.38	0.46			0.13	−0.22
106C	348	0.82	0.67	0.86	−0.07	−0.12	0.32	0.43	0.03	0.42	0.76
107D	404	0.90	1.17	0.83	1.14	0.38	0.56	0.65	0.98	0.56	1.21
108D	381	0.85	0.72	1.14	0.73	0.28	0.39	0.72	0.61	0.89	0.64
791A1	275	1.49	0.81	1.76	1.82	0.85					
791A4	350	1.22	1.56	1.04	2.78	0.99	1.18	1.08	0.60	0.18	0.92
X441	501	0.80			0.71	0.29	0.25		0.15		0.38
Arithmetic mean	419	1.01	0.90	1.13	1.03	0.38	0.44	0.72	0.38	0.43	0.51

transformation. Tables 4–6 give the result of Bartlett's test statistics for homogeneity of variances, skewness and kurtosis for each of the variates measured in the trials studied. A transformation was usually found to increase the homogeneity of variances and reduce skewness and kurtosis. Especially for volume, homogeneity of variance, skewness and kurtosis were improved by a transformation. Branch angle and number of branches seemed to be variates that often deviated from normality or had heterogeneous variances, even after application of a power transformation.

An example on the effects of using power transformation is illustrated in Figs. 1–3 for volume in trial 101E. Figure 1 shows the relationship between log-likelihood and the transformation parameter λ with a 95% confidence interval, Fig. 2 shows that the residuals are more normally distributed after volume has been transformed with a power transformation. The effects on variance, skewness and kurtosis, respectively, in different blocks are shown in Fig. 3. The left side of the diagram shows the results before, and the right side after, transformation. The variance of the transformed values is more homogeneous. The residuals have been standardized to have a mean of 1. Skewness and kurtosis in this example have come closer to zero, and the slope of the lines has become lower. This means that variance, skewness and kurtosis are essentially the same in blocks with low values and in blocks with high values.

Effects of ranking and selection in the simulation study

Half-sib progeny testing

Results of selection for higher trait values are shown in Fig. 4. These diagrams show only the values for variables distributed as y^5, y^{10} and $\exp(y)$ vis-à-vis y, as the less skewed distributions included in the simulations affected the selection only slightly. The selection efficiency decreases and the relative overestimation of the predicted selection gain increases with higher selection intensity. A 10% bias in a predicted selection gain of say + 20% means that the actual gain is predicted as + 22%. The more skewed the distribution is, the more these tendencies increase. In the half-sib case a higher genetic coefficient of variation gave a lower efficiency and a higher overestimation of the genetic gain.

Results when selecting towards low trait values are shown in Fig. 5. Even in this case, the efficiency was lower and the selection gain was overestimated in the skewed distribution. The curves were less steep for higher selection intensity and were less affected by the selection intensity than was selection for higher trait values.

Clone testing

Selection for higher trait values is illustrated in Fig. 6. Contrary to the half-sib situation, the efficiencies were greater and the predicted selection gains became less exaggerated as the genetic coefficient of variation was increased. Selection for lower trait values is illustrated in Fig. 7. Efficiency was lower and the genetic gain was overestimated to greater extent with the low genetic coefficient of variation.

Table 7 shows that normally distributed data give the highest heritability estimates.

Discussion

Common linear procedures are generally considered as robust to slight departures from normality. A trans-

Table 4. Bartlett's test statistic for homogeneity of variances for the variates measured in the different trials. Bartlett's test statistic is approximately χ^2-distributed

Trial code	Number of blocks	H1		H2		INC		DIA		VOL		BD1		BD2		BA1		BA2		BNUM	
		u	t	u	t	u	t	u	t	u	t	u	t	u	t	u	t	u	t	u	t
101A	12	29.0**	29.0**	54.3**	31.4**	14.0	16.6	34.2**	20.0*	251.1**	27.0**	96.5**	49.0**			10.8	10.2			130.5**	48.6**
101C	40	59.4*	36.6					73.0**	34.6	301.1**	33.1	75.4**	46.9			57.5*	45.1			119.9**	84.4**
101D	10	20.5*	13.6					7.2	6.4	22.1**	8.3										
101E	40	69.2**	66.4**					61.6*	61.5*	182.4**	65.0**	57.8*	48.7			30.6	28.1			131.6**	51.4
106C	12	18.0	20.3*	13.1	12.4	12.2	11.5	14.2	5.0	54.1**	6.2	31.4**	22.2*	13.4	10.1	37.1**	31.4**	11.9	12.4	35.4**	32.1**
107D	48	54.2	53.6	44.9	44.6	59.6	59.0	53.9	54.3	86.0**	54.3	56.7	53.0	61.0	57.3	76.1**	76.2**	61.4	61.8	48.8	46.0
108D	36	48.1	41.1	42.4	41.7	50.1*	50.1*	25.1	23.5	49.5	25.8	38.2	30.2	23.5	25.1	53.1*	54.2*	30.9	30.0	52.6*	45.7
791A1	40	48.1	50.2	45.1	39.9	63.6**	55.6*	42.1	29.8	29.3	31.2										
791A4	40	47.2	41.6	35.4	35.7	30.5	30.7	62.5**	43.0	38.7	38.7	64.9**	63.9**	66.5**	65.6**	46.4	45.4	58.4*	53.0	40.8	40.7
X441	32	41.8	42.2					36.2	34.5	91.4**	37.2	54.1**	43.4			39.0	37.6			51.6*	59.4**

Levels of significance: * $P < 0.05$, ** $P < 0.01$
u, untransformed data; t, transformed data using the best suited power transformation

Table 5. Skewness for the variates measured in the different trials. A skewness of zero means that the distribution is symmetric. A positively skewed distribution has a heavy tail to the right

Trial code	H1		H2		INC		DIA		VOL		BD1		BD2		BA1		BA2		BNUM	
	u	t	u	t	u	t	u	t	u	t	u	t	u	t	u	t	u	t	u	t
101A	0.30	0.29	0.12	−0.23	−0.31	−0.22	−0.03	−0.28	0.93	−0.27	0.47	−0.22			0.22	0.00			0.47	−0.24
101C	0.29	−0.10					0.36	−0.11	1.19	−0.12	0.48	−0.05			0.33	0.07			0.94	0.22
101D	−0.32	−0.02					−0.21	−0.07	0.54	−0.15										
101E	−0.06	0.04					0.03	0.04	0.94	0.07	0.33	−0.02			0.30	0.02			0.60	0.14
106C	0.27	−0.01	0.04	−0.08	0.01	−0.10	0.56	0.02	1.27	0.02	0.67	0.10	0.61	0.06	0.58	0.06	0.29	0.02	0.24	−0.02
107D	−0.08	−0.01	0.01	−0.02	0.02	−0.03	−0.15	−0.11	0.25	−0.12	0.25	0.03	0.20	−0.01	0.22	0.20	0.30	0.15	−0.09	0.02
108D	−0.02	−0.20	−0.10	−0.15	−0.17	−0.12	−0.02	0.13	0.50	−0.13	0.19	−0.15	0.10	−0.08	0.34	0.18	0.26	0.21	0.18	−0.08
791A1	0.20	0.11	−0.06	0.12	−0.31	−0.04	−0.31	0.06	0.24	0.12										
791A4	−0.29	−0.04	−0.23	−0.16	−0.25	−0.24	−0.49	−0.03	−0.01	−0.01	−0.04	0.04	−0.09	−0.04	0.32	0.15	0.33	−0.03	−0.04	−0.08
X441	0.16	0.09					0.11	−0.03	0.73	0.00	0.29	−0.07			0.39	0.07			0.34	−0.02
Arithmetic mean	0.05	0.02	−0.04	−0.09	−0.17	−0.13	0.02	−0.04	0.66	−0.06	0.33	−0.04	0.21	−0.02	0.34	0.09	0.30	0.09	0.33	−0.01

u, untransformed data; t, transformed data using the best suited power transformation

Table 6. Kurtosis for the variates measured in the different trials. A kurtosis of zero means that the peakedness is the same as for the normal distribution. Peaked distributions show positive kurtosis and flattopped distributions show negative kurtosis

Trial code	H1		H2		INC		DIA		VOL		BD1		BD2		BA1		BA2		BNUM	
	u	t	u	t	u	t	u	t	u	t	u	t	u	t	u	t	u	t	u	t
101A	0.00	0.02	−0.36	−0.22	0.06	−0.06	0.04	0.10	0.98	−0.22	0.35	0.41			−0.23	−0.24			1.20	1.39
101C	−0.11	−0.27					−0.17	−0.33	1.56	−0.37	0.57	0.14			0.12	−0.02			2.64	1.64
101D	0.15	−0.14					−0.21	−0.34	−0.06	−0.33										
101E	−0.02	−0.06					−0.13	−0.14	0.71	−0.20	0.49	0.29			−0.17	−0.31			0.86	0.35
106C	−0.15	−0.21	−0.43	−0.43	−0.43	−0.41	−0.16	−0.55	1.54	−0.66	0.54	0.12	0.70	0.14	0.45	0.08	0.01	−0.15	1.02	1.03
107D	0.11	0.10	−0.01	0.00	0.00	−0.01	−0.23	−0.24	−0.36	−0.39	0.40	−0.35	−0.25	−0.29	0.98	0.98	−0.18	−0.21	−0.04	−0.10
108D	0.13	0.21	−0.06	−0.04	−0.22	−0.25	−0.30	−0.30	−0.03	−0.38	−0.01	0.01	−0.04	−0.03	−0.02	−0.12	0.03	0.01	0.69	0.67
791A1	−0.18	−0.20	−0.16	−0.24	0.25	0.07	0.36	0.10	0.00	−0.01										
791A4	0.15	−0.10	0.37	0.32	0.41	0.40	0.07	−0.41	−0.28	−0.28	0.21	0.20	−0.42	−0.42	0.30	0.23	0.28	0.22	0.42	0.44
X441	0.03	0.00					−0.04	−0.07	0.70	−0.04	−0.16	−0.18			0.14	−0.12			0.85	0.61
Arithmetic mean	0.01	−0.07	−0.11	−0.10	0.01	−0.04	−0.08	−0.22	0.48	−0.29	0.30	0.08	0.00	−0.15	0.20	0.06	0.04	−0.03	0.96	0.75

u, untransformed data; t, transformed data using the best suited power transformation

Fig. 1. Plot of likelihood function L versus the Box-Cox transformation parameter λ for volume in trial 101E Lekvattnet. The 95% confidence limits (λ_1, λ_2) are also shown

Fig. 2a, b. Example of frequency distribution of the residuals of volume before (**a**) and after (**b**) transformation in trial 101E Lekvattnet

formation solely to achieve normality can be chosen within a fairly wide interval (Box and Cox 1964). The requirement of constant error variance exercises greater influence on the choice of a transformation than does the requirement of normality (i.e. Ibe and Hill 1988). Therefore, Ibe and Hill (1988), in their example with poultry egg production, first used a power transformation to produce a more normally distributed variable, then they used a scalar to modify the transformation parameter to produce homogeneity of variances. Box and Cox (1964) suggested a similar procedure by first applying a transformation by prior reasoning and only thereafter considered necessary modifications. The empirical power transformation in

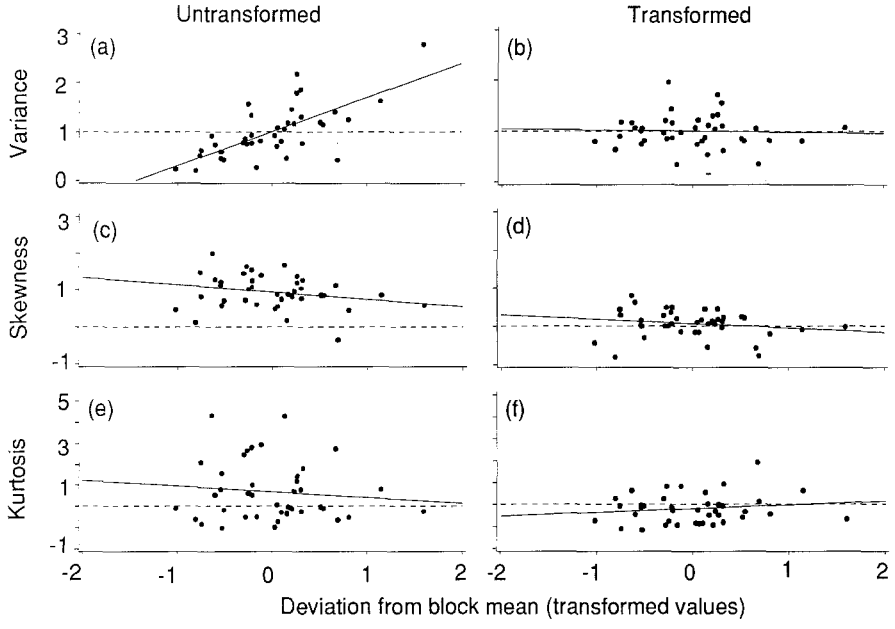

Fig. 3a–f. Plots for checking homogeneity of variance, skewness and kurtosis of volume, exemplified for trial 101E Lekvattnet. Each *dot* represents the value of one block. The block mean values are expressed as deviations from the block mean of the transformed values

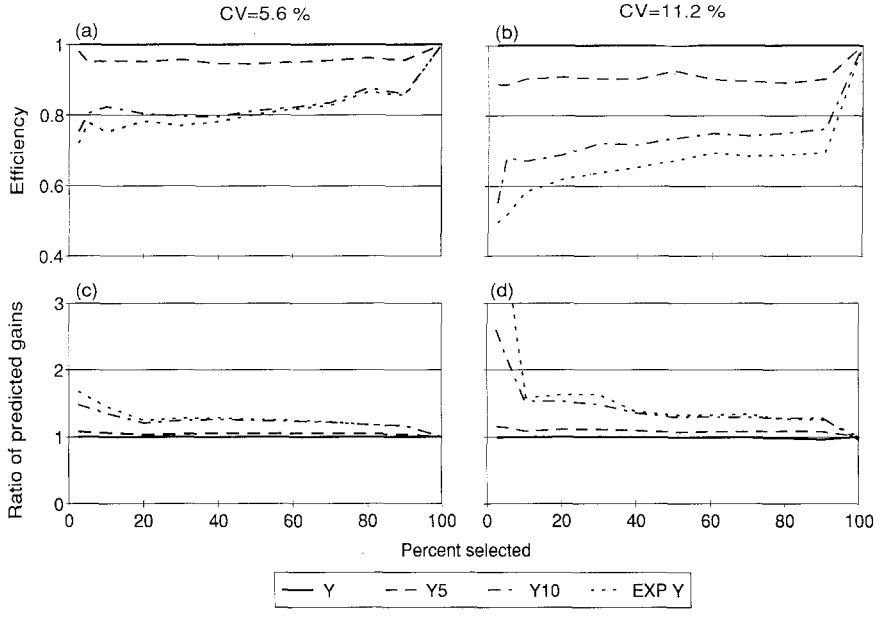

Fig. 4a–d. Obtained genetic gain, efficiency and bias in predicted selection gain for simulated half-sib data within different distributions. Selections were made towards higher trait values. Y = normally distributed y, $Y5 = y^5$, $Y10 = y^{10}$, $EXP\ Y = \exp(y)$

our study was often successful in improving normality and homogeneity of variances at the same time. The method suggested by Sokal and Rohlf (1981), model (8) above, which is a composite of L and Bartlett's homogeneity of variances test statistic, did not improve the situation significantly compared with a pure power transformation. Even if, for example, $\sqrt{y}$ is found to be the best transformation from the formal analysis, it could be more convenient to work with the logarithm of y, while there are arguments of ease of interpretation (Box and Cox 1964).

According to Ibe and Hill (1988) there is some evidence that transformed data yield higher heritability estimates than untransformed data. A higher heritability gives rise to higher expected selection responses when testing efforts are constant. The simulation study done here confirms this, although differences were small unless the distribution was heavily non-normal (Table 7).

A disadvantage of power transformations is that there is no genetic or other biological justification for the transformation, except that polygenic effects are

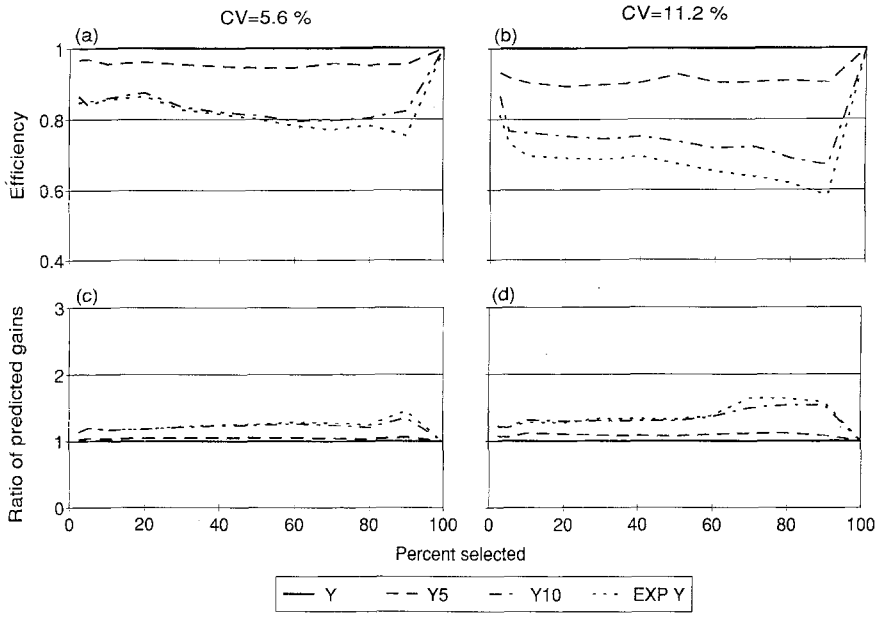

Fig. 5a–d. Obtained genetic gain, efficiency and bias in predicted selection gain for simulated half-sib data with different distributions. Selections were made towards lower trait values. Y = normally distributed y, $Y5 = y^5$, $Y10 = y^{10}$, $EXP\ Y = \exp(y)$

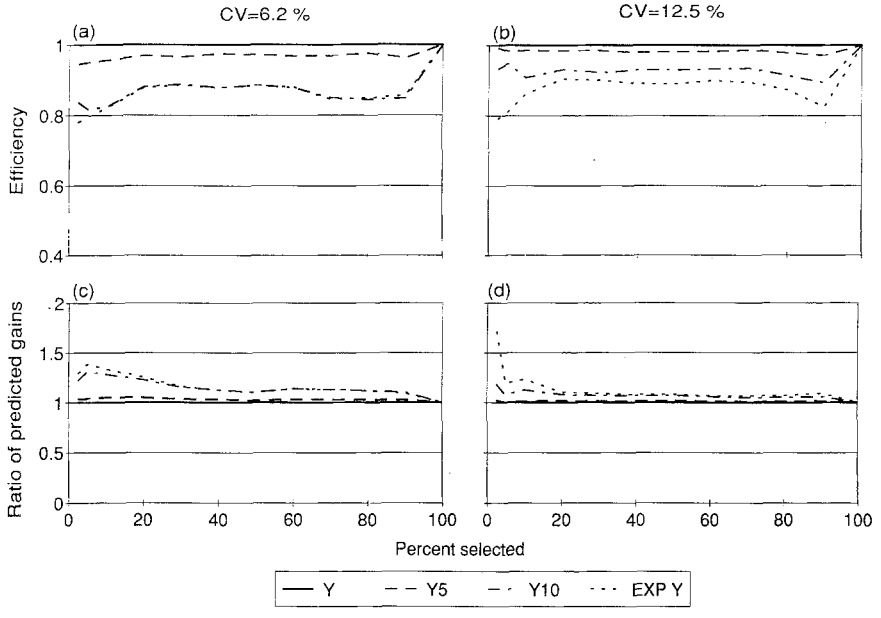

Fig. 6a–d. Obtained genetic gain, efficiency and bias in predicted selection gain for clone testing on simulated data with different distributions. Selections were made towards higher trait values. Y = normally distributed y, $Y5 = y^5$, $Y10 = y^{10}$, $EXP\ Y = \exp(y)$

expected to be normally distributed due to theoretical considerations. Most variates are measured in easily understood units, such as, cm, m^3 etc., but on the transformed scale it is more difficult to understand the meaning of a value. Therefore, a retransformation of genetic values to the original scales is often desired. For some scales the backward transformation is biased and should be adjusted. Table 8 lists adjustment terms that could be added to the retransformed values when using transformations from the power family. These approximate adjustments are derived from Taylor series expansions and are one-half the product of the second derivative of the backward transformation function and the population variance of the transformed variable, as suggested by Kung (1988) based on Kruskal (1978).

The simulation study revealed two problems associated with skewed distributions in genetic evaluation and selection (1) lower efficiency, which means that incorrect parents or clones are selected, and (2) overestimation of predicted selection gain.

In the simulation study the selection efficiency proved to be dependent on the degree of genetic variability. With half-sib data, in contrast to clone data, a

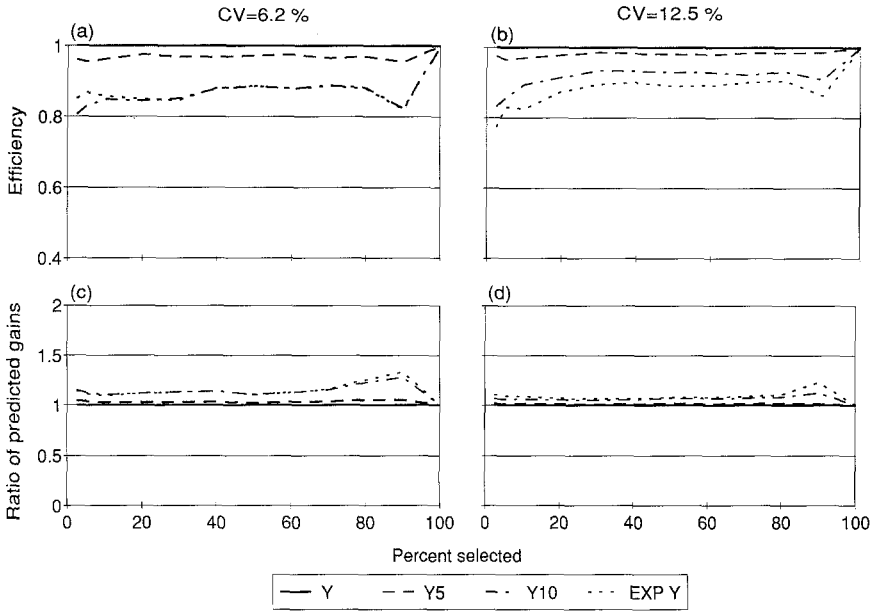

Fig. 7a–d. Obtained genetic gain, efficiency and bias in predicted selection gain for clone testing on simulated data with different distributions. Selections were made towards lower trait values. Y = normally distributed y, $Y5 = y^5$, $Y10 = y^{10}$, $EXP\ Y = \exp(y)$

higher genetic coefficient of variation reduces the efficiency. There is no obvious explanation for that. In both cases with constant heritabilities, increased coefficients of variation produced proportionately similar changes in variation between and within genetic entries. The only clear difference between these two cases is the distribution of variation between and within genetic entries relative to large block effects. The relative increase in the residuals is larger in the half-sib analyses. It was beyond the scope of this investigation to study this aspect in depth, however. The simulations will probably exaggerate the effects of the different distributions to some extent. In the simulations with unrestricted random numbers, extreme variate values will presumably appear more frequently than they do in nature. For example, there are biological limits for height or volume of a tree that were not implemented in the simulation model.

With reference to the datasets available in this study, there seems to be no general need for transformations for height, height increment and diameter values in tree breeding data. The distributions of volume, branch diameter, branch angle and number of branches often showed a positive skewness. Selection for branch diameter and number of branches is usually made towards lower values, which means selection from the left tail of the distribution. Even though there is a bias in selection gain, these selections are scarcely affected in absolute values, as the real genetic gain is small. For some variates, such as the number of branches, it is difficult to remove skewness, kurtosis and heterogeneity of variances, even with power transformation; the power transformation may not be the appropriate one for these variates. If needed at all, a

Table 7. Estimated heritabilities in the different distributions of simulated data: narrow-sense heritability (h^2) for half-sib data and broad-sense heritability (H^2) for clone data. Expected values in the normal distribution are $h^2 = 0.20$ and $H^2 = 0.25$

Distribution[a]	Assumed genetic coefficient of variation (CV)			
	h^2		H^2	
	5.6	11.2	6.2	12.5
$y^{0.5}$	0.185	0.192	0.260	0.240
y (i.e. normal distribution)	0.185	0.196	0.261	0.243
y^2	0.180	0.191	0.256	0.235
y^5	0.140	0.131	0.203	0.169
y^{10}	0.065	0.040	0.102	0.054
$\exp(y)$	0.056	0.021	0.092	0.009

[a] y normally distributed

Table 8. Adjustment for bias in backward transformation. The adjustment should be added to the backward transformed value (form Kung 1986)

Classification	Transformation		Adjustment
	Forward	Backward	
Square root	$\bar{y} = \bar{x}^{0.5}$	$\bar{x} = \bar{y}^2$	$\hat{\sigma}_y^{2a}$
Reciprocal	$\bar{y} = 1/\bar{x}$	$\bar{x} = 1/\bar{y}$	$\hat{\sigma}_y^2/\bar{y}^3$
Logarithm	$\bar{y} = \ln(\bar{x})$	$\bar{x} = \exp(\bar{y})$	$\hat{\sigma}_y^2 \exp(\bar{y})/2$
Exponential	$\bar{y} = \exp(\bar{x})$	$\bar{x} = \ln(\bar{y})$	$-\hat{\sigma}_y^2/2\bar{y}^2$
Quadratic	$\bar{y} = \bar{x}^2$	$\bar{x} = \bar{y}^{0.5}$	$-\hat{\sigma}_y^2/8\bar{y}^{1.5}$
Power	$\bar{y} = \bar{x}^c$	$\bar{x} = \bar{y}^{1/c}$	$\hat{\sigma}_y^2 \bar{y}^{1/c}(1-c)/2c^2\bar{y}^2$

[a] $\hat{\sigma}_y^2$ is the population variance of transformed values

simple normal scores approach (Danell 1988) may be theoretically more suitable for categorical traits such as number of branches.

The conclusion that was drawn from the simulations in this study is that a heavily skewed distribution both impairs the selection efficiency and exaggerates the selection gain expectations. Our recommendation in the context of forest tree breeding data is that, prior to genetic evaluations, one should check the distribution of the variate measured. If the distribution is seriously skewed, a transformation might be worthwhile. It does not seem possible, with reference to the available empirical datasets in this study, to give any general guidance about which variate should be transformed, but in most cases there seems to be no need for transformations.

Acknowledgements. The authors express their thanks to Dr. G. Ekbohm and Prof. G. Eriksson for valuable comments on the manuscript.

References

Andersson S-O (1954) Funktioner och tabeller för kubering av småträd. Meddelanden från Statens skogsforskningsinstitut 44(12). Stockholm

Bartlett MS (1947) The use of transformations. Biometrics 3:39–52

Box GEP, Cox DR (1964) An analysis of transformations. J R Stat Soc B26:211–243

Danell Ö (1988) Theoretical aspects in the estimation of breeding values for all-or-none traits. Report 77, Department of Animal Breeding and Genetics, Swedish University of Agricultural Sciences, Uppsala

Henderson CR (1975) Best linear unbiased estimation and prediction under a selection model. Biometrics 31:423–447

Ibe SN, Hill WG (1988) Transformation of poultry egg production data to improve normality, homoscedasticity and linearity of genotypic regression. J Anim Breed Genet 105:231–240

Kruskal JB (1978) Transformation of data. In: International encyclopedia of statistics:1044–1055. Macmillan Publ. London

Kung FH (1988) Application of data transformation in forest genetics. Silvae Genet 37:45–49

Näslund M (1947) Tabeller och funktioner för kubering av stående träd. Meddelanden från Statens skogsforskningsinstitut 36(3). Stockholm

Neyman J, Scott EL (1960) Correction for bias introduced by a transformation variables. Ann Mat Stat 31:643–665

SAS Institute Inc (1990) SAS/STAT® user's guide, version 6, 4th edn. SAS Institute Inc, Cary, N.C.

Sokal RR, Rohlf FJ (1981) Biometry, 2nd edn. WH Freeman and Co. San Francisco