Prof. Dr. Helmut Schmidt-Vogt – 70 Jahre

Am 8. 1. 1918 in NO-Bayern geboren, teilte der Jubilar das Schicksal vieler Zeitgenossen: Das Dritte Reich, die Kriegs- und Nachkriegszeit zeichnete seine Jugend, Ausbildung und erste Berufsausübung. Unmittelbar nach dem Abitur wurde er zum Reichsarbeits- und dann zum Wehrdienst eingezogen. Schwerkriegsbeschädigt entließ man ihn 1943 zum Studium an die Universität München. 1945 bestand er bereits das Diplom-Examen, 1947 die große, forstliche Staatsprüfung bei der bayerischen Staatsforstverwaltung, nachdem man ihm in der Referendarzeit Aufgaben übertragen hatte, wie man sie heute nur „gestandenen" Amtsleitern übertragen würde.

Umfassende berufliche Erfahrung erwarb er sich in den Folgejahren an einer Oberforstdirektion, der Ministerialforstabteilung in München und vor allem in Teisendorf, einem der vielseitigsten bayerischen Forstämter, das noch eine Reihe von Nebenbetrieben, wie eine überregional bedeutsame Baumschule, zu betreuen hatte.

Neben der hier nötigen Aufbauleistung gelang es Professor Schmidt-Vogt noch, sich intensiv in wissenschaftliche Problem des Waldbaus und besonders der Pflanzenzucht einzuarbeiten. Seine Promotion 1950, Habilitation 1962 und zahlreiche Veröffentlichungen belegen dies.

1964 folgte er einem Ruf auf den Lehrstuhl für Waldbau an der Universität Freiburg und ging unverzüglich an den Aufbau eines experimentell ausgerichteten Instituts und zahlreicher Forschungsprojekte, die reichen Niederschlag in Form von Veröffentlichungen fanden.

Bald darauf entschloß er sich, eine umfassende Monographie über unsere wichtigste Baumart, die Fichte, zu schreiben. Auf Reisen um die ganze Nordhalbkugel studierte er alle bedeutsamen Fichten-Waldtypen vor Ort und verpflichtete dank seiner geduldigen, verbindlichen, überzeugenden und beharrlichen Art eine große Zahl in- und ausländischer Wissenschaftler zur Mitarbeit an diesem gigantischen Unterfangen. Zwei hervorragend gestaltete und sorgfältig erarbeitete Bände der Monographie liegen vor, der eine bereits in zweiter Auflage. Zwei weitere stehen kurz vor dem Abschluß.

Das alles brachte mancherlei übernationale Aufgaben und Ehrungen mit sich, war aber nur möglich durch die ausdauernde Unterstützung seiner Frau und zugleich erkauft mit hartem persönlichem Einsatz und Entbehrungen.

Seine früheren Mitarbeiter ebenso wie seine Kollegen wünschen ihm für die nächsten Jahre Kraft und Zähigkeit, um sein Lebenswerk zu vollenden und noch einige jener Vorhaben wieder aufgreifen zu können, die lange hatten zurückstehen müssen.

J. Huss

Application of Data Transformation in Forest Genetics

By F. H. Kung[1])

(Received 16th May 1986)

Summary

Stabilizing variance, improving additivity and simplifying relationships are possible uses of a suitable data transformation. However, transformed models may be inferior in terms of difficulties in mental comprehension, bias in backward transformation, change in error minimization, decrease in F-ratio and contradiction to reality. The benefits and the drawbacks were illustrated by a genetic study of white ash. Theoretical considerations and empirical methods for selecting transformations are presented with examples.

Key words: Homoscedasticity, power families, white ash.

Zusammenfassung

Durch die Wahl geeigneter Umformungen, werden eine stabilere Varianz, eine verbesserte Additivität des Datenmaterials, sowie die Herstellung einfacher, linearer Beziehungen erreicht. Mögliche Schwächen des Transformationsmodells liegen im schwierigen Verständnis dieser Umformungen, in Fehlern bei der Rückumformung, in Veränderungen in der Fehlerminimierung, in der Abnahme im F-Verhältnis und in einem gewissen Widerspruch zur Realität. Die Vor- und Nachteile des Modells werden anhand einer genetischen Untersuchung an *Fraxinus americana* L. veranschaulicht. Theoretische Erwägungen und empirische Methoden zur Auswahl der Umformungen werden an Beispielen dargestellt.

Introduction

Before the age of computers, data transformation by hand calculation or through tables was laborious and error-prone. With the advance and ready availabilty of computers, many transformation subroutines are accessible through programming. It is as easy to use the transformed data as the original observations in statistical data analyze. The use of transformation in forest genetics literature has increased but still relatively few authors consider such alternatives. When plot means are used in the analysis of variance, we may take the "Central Limit Theorem" in mathematical statistics for granted and may be content with the thought that the required normal assumptions are automatically met. To illustrate, the distribution of individual tree diameters may not be normal, but the distribution of plot means of the diameters can be considered as normal when the plot size is large. Therefore, statistical inference concerning provenance differences and plantation differences should be valid when the plot means is used as observation. However, when the plot size is small or when blind faith in the "Central Limit Theorem" is insufficient, one should always study the residual errors and try to find a suitable transformation so that the normal assumption may be met. The objectives of this paper are (1) to demonstrate the benefits of data transformation, (2) to illustrate the methods of choosing a suitable transformation, and (3) to discuss the practical difficulties with transformation.

[1]) Professor, Department of Forestry, Southern Illinois University, Carbondale, IL 62901, USA.

Needs of Data Transformation

Most parametric statistical inferences are dependent on its underlying distributions. For example, mutation is a rare event and theoretically the occurrence of mutation should be considered as a Poisson variate, and that the statistical inference should be based on the Poisson distribution and not the normal distribution. However, if an experiment for mutation were set up as a factorial design, where analysis of variance was an appropriate test of treatment means, the square root transformation could be used (Bartlett, 1936). Another example can be given by the germination percentage. Percentage is a binomial variate and we should use statistical methods based on that distribution. However, if we preferred not to use the test which is based on the binomial distribution but rather preferred to use the normal approximation method, we could transform the germination percentage to arcsine and proceed as if the data were normal variates (Bartlett, 1937; Cochran, 1940). Thus, in order to meet the required assumptions for a given statistical test, transformation is often appropriate.

Analysis of variance also requires additivity. Transformation may be useful in improving additivity and reducing interaction. Interaction is the failure of genetic entities to maintain the same relative ranks and level of differences when tested in different environments (Snyder, 1972). If the interaction were due to the failue to maitain the same ranks, monotonic transformation would not reverse the trend. But, if the interaction were due to the failure to maintain the same level of differences, a suitable transformation in scale of measurement could reduce or eliminate the interaction. For example, in a white ash genetic experiment with three factors (year, provenance, and plantation), the plantation by year interaction became null and the variance component for provenance by plantation reduced from 8.5 percent to 6 percent when the height growth data were transformed to the logarithmic scale (Kung and Clausen, 1983). Thus, for analysis of variance in this experiment, growth expressed in the logarithmic scale is more suitable (more additive) than the original linear scale. In other words, growth measured in percentage seems to be more additive than growth in inches or cm.

Transformation changes the relationship between variables. Because a straight line relationship is much easier to comprehend and interpolate than a curvilinear relationship, transformation may be useful in simplifying data relationships. For example, by taking a natural logarithmic transformation of the age ratio (year of mature age/year of juvenile age), Lambeth (1980) was able to express the age-age correlation with a simple linear regression line. As with straight line relationships between variates, straight lines of confidence band are also easier to interpret than curvilinear bands. By taking a square root transformation, the acceptance region of a genetic segregation ratio can be approximated by a straight band. For example, the approximate 95 percent acceptance region for the 3:1 segregation ratio can be outlined by first drawing two circles with the center at (30,10) and at (300,100) and with the radius equal to the length of 2 standard-error-full-scale on a binomial paper, then two parallel tangents are drawn to these two circles (*Figure 1*). The length of 2 standard-error-full-scale is chosen to approximate the 95 percent confidence level. Any data point from an experiment falling outside the acceptance region may serve as evidence to reject the null hypothesis of 3:1 segregation ratio. The acceptance regions are also useful for sample size determination. For example, after the acceptance regions for the 3:1 and the 15:1 ratios were plotted, Kung (1975) reported that a minimum of 55 samples were needed to accept one and reject the other hypothesis. If the sample sizes were less than 55, it would be possbile that both ratios be accepted. Such solutions would be difficult to obtain if the data were not plotted on the square root transformation.

Choosing a Suitable Transformation

Observations obtained from a forest genetic experiment may not follow any standard theoretical distribution. Most diameter and volume distributions have a long tail to the right and are not symmetrical. The assumption of homogeneity of variance is also difficult to meet because fast growing families are more variable than slow growing families. Nevertheless, if a suitable transformation were chosen, it usually normalizes the data and at the same time stabilizes the variance.

To choose a suitable transformation for analysis of variance we should plot the standard deviation against the mean for each family on a graph. If it shows a straight line relationship (slope is not important), the logarithmic transformation would be the choice. Otherwise, one should try to plot the standard deviation against the square root or the square of the family mean. If the use of the square root of family mean produced a straight line, then the data should be transformed by its square root. But if the square of the family mean would fit better, then one should choose the reciprocal transformation. Finally, if none of the above three simple transformations would work, we should fit the standard deviations and family means into the following regression:

$$\ln(\text{standard deviation}) = a + b \ln(\text{family mean}).$$

Once the coefficient for b is obtained, each observation y is then transformed to $(y^{(1-b)}-1)/(1-b)$. The transformed data now can be used for statistical analysis (Box and Cox, 1964; Box, Hunter, and Hunter, 1978). In analysis of variance because the F-ratio remain invariant under changes of location and scale, use of y^b and $(y^{(1-b)} - 1)/(1-b)$ leads to identical tests of significance. In this case, we can simply transform y to the b power (Hinz and Eagles, 1976).

Let me illustrate with a study of interaction in a white ash involving 19 provenances in Louisiana, Illinois, Ohio and Wisconsin (Kung and Clausen, 1983). The variance of error within the Illinois plantation (3340) was 13 times as large as that in the Wisconsin plantation (254). But after the logarithmic transformation of height was used, no significant difference was found between the variance in Illinois (0.077) and in Wisconsin (0.089).

To choose a suitable transformation for regression analysis, the "Matchacurve" (Jensen and Homeyer, 1971) is a graphical method to determine the power transformation for the input variable. The data points are first scaled and plotted to a standard format of 7.5 inches (190 mm) horizontally and 5.0 inches (127 mm) vertically, and then they are compared to the wide array of standard curves presented in that publication. If a standard curve can be found to match the data points, then the algebraic transform specified for the standard can be applied to the variable. For example, when we plotted the variance component for family versus age in a genetic study of black walnut (Rink, 1984) the data points matched closely with the standard curve of power 2.5 (*Figure 2*). Therefore, the regression line

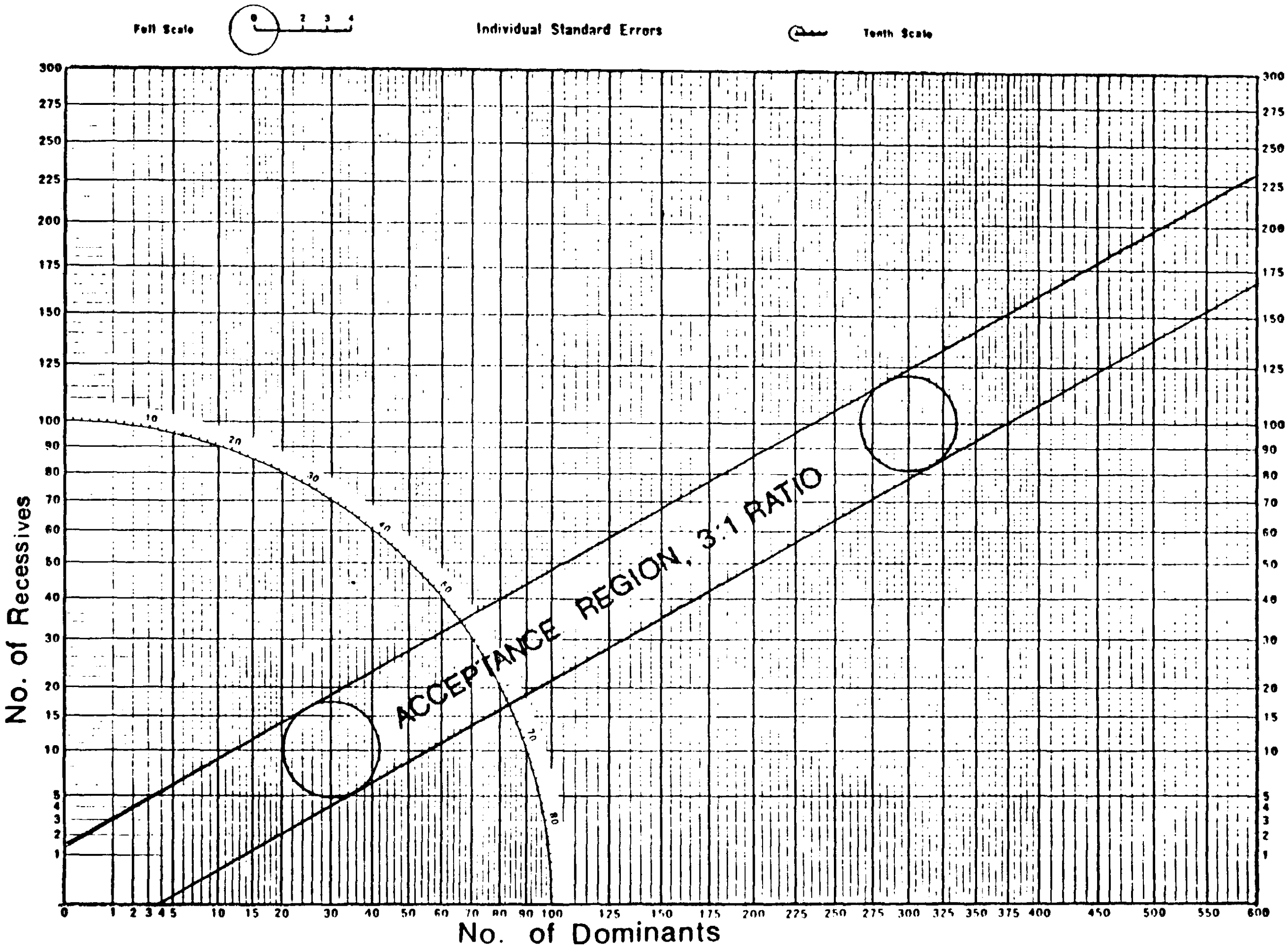

Figure 1. — Plotting acceptance region on a binomial paper, with the square root transformation.

could be expressed as family variance component = a + b(age)$^{2.5}$.

Another rapid method for choosing a transformation was suggested by Dolby (1963). The method involves selection of four equally spaced x variables. They are fitted to a third order differential equation and the constant of this equation provides a convenient index for the transformation. More specifically, let

$$A = 0.5 \left\{ 1 + \frac{(D3 + D2)}{(D3 - D2)} - \frac{(D2 + D1)}{(D2 - D1)} \right\}$$

where: D3 = (Maximum of x) — (x at 66 percentile)
D2 = (x at 66 percentile) — (x at 33 percentile)
D1 = (x at 33 percentile) — (Minimum of x)

The power to be raised is (A + 0.5)/(A — 0.5). Some special cases for the value of A and its corresponding transformation are listed below:

A	Transformation
1.5	Quadratic
1.0	Cubic
0.5	Exponential
0	Reciprocal
−0.5	Logarithm
−1.5	Square Root

Using the same data from Rink (1984), the family variance components at age 1, 4, 7 and 10 years were respectively 0.00451, 0.01223, 0.05014 and 0.12945. The index value of A is calculated at 1.16 and the power to be raised for the age variable is 2.51, similar to the "Matchacurve" result. Furthermore, when his 8 original data were used, the regression: Family variance component = 0.0008 + 0.0004(Age)$^{2.5}$ was statistically significant with R-square of 0.997.

Difficulties with Transformation

Although suitable data transformations may stabilize variance, improve normality, simplify relationships, and allow subsequent stages of analysis to be more accurate and more revealing, they may simultaneously create some difficulties and confusion. In essence there are some short-

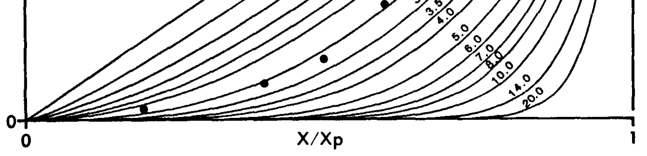

Figure 2. — Plotting data points on "Matchacurve" to choose a power transformation. The power of 2.5 seems to be suitable for the data presented.

comings in terms of mental comprehension, backward transformation, error minimization, goodness of fit, and practical application.

1. Mental Comprehension

Most observations are recorded in units commonly encountered in our daily life and we can quickly grasp the magnitude of the attribute represented by the original scale. For example, we can easily see the contrast between a 80 percent and a 20 percent germination rate. The difference is 60 percent and the former is four times greater than the latter. On the other hand, our minds are seldom familiar with the transformed units. Sometimes, the transformed unit may be too abstract to understand. To illustrate, if the above two germination rates were represented in arcsines, most of us could not intuitively grasp the meaning of the difference between 64 and 27 degrees. In some cases the transformed dimension may be meaningful: the square transformation of diameter can be considered as basal area, the reciprocal of age involves the function of the mean annual growth, and the reciprocal of area is a function of density.

2. Backward Transformation

To refer back to the original scale and to faciliate interpretation, backward transformation is needed. But the backward transformation is biased (Neyman and Scott, 1960), because "the transform of the expected value does not equal the expected value of the transform" (Kruskal, 1978). To illustrate, the mean of three values 0.1, 0.2 and 0.6 is 0.3. The mean of their square root transformations is 0.5127. If we took the backward transformation, the square of 0.5127 is 0. 2628, which is smaller than the original mean of 0.3.

One simple improvement for the bias in the backward transformation is to add an adjustment term to it. The approximate adjustment is one half the product of the second derivate of the backward function and the population variance of the transformed variable (Kruskal, 1978). The adjustment terms for the 6 power families are listed in *Table 1*. Because the approximate adjustment is based on a Taylor expansion through the quadratic term, it is exact only for the square root transformation. Using the given example (x = 0.1, 0.2, and 0.6), we can see in *Table 1* that the adjustments are quite sufficient for logarithmic, exponential, and quadratic transformations. The improvement in the reciprocal transformation is substantial but the transform of the expected value is still lower than 0.3 by about 17 percent. As to the power transformation, the closer it is to the power of 2, the greater is the improvement in accuracy with adjustment than the straight backward transformation without adjustment·

Other general methods for bias correction in backward transformations are available (Neyman and Scott, 1960). For logarithmic transformation, Wiant and Harner (1979) used the error variance after fitting a regression model rather than using the population variance of the transformed variable before fitting. Their method seems to be more specific to the model than the general approximate adjustment and therefore should be more precise.

3. Error Minimization

When the independent variables are transformed in the least squares fitting, the sum squares for error between the observed and the fitted values on the original scale is minimized as usual· But when the dependent variable is transformed, the least squares solution is applied to the transformed and not to the original scale. For example in fitting the logistic growth curve, y = a/{1 + exp(b-cx)}, one can use non-linear regressing or transformed linear regression methods. In the transformed linear regression method, the growth function is transformed into ln{(y-a)/a} = b-cx. The upper limit of carrying capacity, a, is first initialized and the quantity ln{(y-a)/a} can be regressed on x. By trial and error one can find a good estimatte of "a" which will have the least squared error. However, the least squares solution in the linear transformation minimizes the squared distance from ln{(y-a)/a} to the regression line, in contrast to the non-linear regression with the original model which minimizes the squared distance from observed y to the curve. This indicates that if we want a good prediction on the original scale, and not on the transformed scale, we should not transform the dependent variable. The transformation of the dependent variable should be done only if we would like to minimize error on the transformed scale. For example, to study the genetic by environment interaction in white ash the logarithmic transformation on height growth was used to test the hypothesis of no differences in growth proportion (Kung and Clausen, 1983). Since the additive error in the logarithmic regression becomes multiplicative error after backward transformation, minimizing the relative error

Table 1. — Adjustment for bias in backward transformation.

Classification	Transformation		Adjustment	Example *				
	Forward	Backward		$\bar{y}$	Vy	$\bar{x}$	Adj	$\hat{x}$
Square Root	$y = x^{.5}$	$x = y^2$	Vy	.5127	.0372	.2628	.0372	.3000
Reciprocal	$y = 1/x$	$x = 1/y$	Vy/y^3	5.5556	11.7284	.1800	.0684	.2484
Logarithm	$y = \ln(x)$	$x = \exp(y)$	$Vy \exp(y)/2$	-1.4743	.5442	.2289	.0623	.2912
Exponential	$y = \exp(x)$	$x = \ln(y)$	$-Vy/2y^2$	1.3829	.0987	.3242	-.0258	.2984
Quadratic	$y = x^2$	$x = y^{.5}$	$-Vy/8y^{1.5}$	.1367	.0251	.3697	-.0621	.3076
Power**	$y = x^c$	$x = y^{1/c}$	$Vy\ y^{1/c}(1-c)/2c^2y^2$	.1953	.0369	.3366	-.0362	.3004

*) Example is given for 3 observations of x (0.1, 0.2 and 0.6) and the symbols are:
$\bar{y}$ = mean of tranformed values
Vy = variance of transformed values
$\bar{x}$ = backward transformation of $\bar{y}$
Adj = adjustment to be added to $\bar{x}$
$\hat{x}$ = adjusted backward transformation of $\bar{y}$
**) C = 1.5 was used for the numeric example.

Table 2. — Comparison of R-square and F-ratio for the response surface model with and without the use of arcsine transformation.

Regression	Arcsine Transformation			
	Before		After	
	R-square	F-ratio	R-square	F-ratio
Linear	0.1548	14.67	0.1148	9.68
Quadratic	0.3118	29.55	0.3618	30.52
Crossproduct	0.1641	31.12	0.1084	18.28
Total	0.6307	23.91	0.5850	19.74

rather than the absolute error in the original scale is appropriate in testing that hypothesis.

4. Goodness of Fit

It is possible that the r^2 for the model after transformation becomes smaller than the original one. For survival of white ash (Kung and Clausen, 1984), the R-square and the F-ratio were lower for the arcsine transformation than for the original survival percent (*Table 2*). Thus, theoretically, arcsine transformation should be used for percentage, but empirically, the data fit the model better on the percentage scale than the arcsine scale. Under the circumstances, one must evaluate the trade-off between (1) meeting the assumptions, and (2) obtaining a good fit, so that a useful model can be selected.

5. Practical Application

Sometimes in tree improvement we are interested in predicting response value y for a given fixed value of the regression variable x. The predictions vary with different models but the range of x and y must be reasonable· For example, when we fitted a second degree response surface model for height growth in white ash, using the latitudes of the seed source and the latitude of the planting location as independent variables we found that the best planting location should be at 37.8 degrees while the best seed sources should be at 34.0 degrees north latitude (Kung and Clausen, 1984). Although these seem reasonable, we were searching for alternative models because in the original scale the residual variance within the Wisconsin plantation was smaller than that in the orthers. When we used the natural logarithmic transformation for height, the error variances did become stable, but the resulting response model has a center of superior seed sources at 58 degree south latitude. We know very well that the natural range for white ash does not extend below 28 degrees north latitude, in this case the model with logarithmic transformation is impractical. Forced to choose between a practical model which might not meet the assumption and an impractical model which would meet the requirements, I would rather sacrifice statistical rigor in favor of practicality.

Conclusion

Data transformation is beneficial in simplifying relationships, stabilizing variance and improving normality. Frequently, a single transformation may achieve two or all three objectives at once. Computational costs for transformation if the computers are available is negligible, but one must also consider that the transformed model should be understandable and practical. Also, it should minimize the right error terms and maximize the goodness of fit.

Acknowledgements

The author would like to thank Drs. Calvin F. Bey, David T. Funk, F. Thomas Ledig, Barbara G. McCutchan and Anthony E. Squillace for reviewing the manuscript and giving many editorial improvements, and to thank Mr. H.-J. Spors for the German translation.

Literature Cited

Bartlett, M. S.: The square root transformation in the analysis of variance. Suppl. Jour. Roy. Stat. Soc. 3: 38—78 (1936). —— Bartlett, M. S.: Some examples of statistical methods or research in agriculture and applied biology. Suppl. Jour. Roy. Stat. Soc. 4: 137—183 (1937). —— Box, G. E. P. and Cox, D. R.: An analysis of transformations. J. Roy. Stat. Soc., Ser. B 26: 211—243 (1964). —— Box, G. E. P., Hunter, W. G. and Hunter, J. S.: Statistics for Experimenters: An Introduction to Design, Data Analysis and Model Building. Wiley Interscience, N.Y., N.Y. 653 p. (1978). —— Cochran, W. G.: The analysis of variance when experimental errors follow the Poisson or binomial laws. Ann. Math. Stat. 11: 335—347 (1940). —— Dolby, J. L.: A quick method for choosing a transformation. Techometrics 5: 317—325 (1963). —— Hinz, P. N. and Eagles, H. A.: Estimation of a transformation for the analysis of some agronomic and genetic experiments. Crop Science 16: 280—283 (1976). —— Jensen, C. E. and Homeyer, J. W.: Matchacurve — 2 for algebraic transforms to describe curves of the class x^n. USDA For. Serv. Res. Pap. INT-106. 39 p. (1971). —— Kruskal, J. B.: Transformation of data. In: International Encyclopedia of Statistics. Macmillan Pub. Co. 1350 p. (1978). —— Kung, F. H.: A Handbook of Graphical Solutions to Forest Biometric Problems. Dept. of For. Pub. No. 12. Southern Illinois Univ. 89 p. (1975). —— Kung, F. H. and Clausen, K. E.: Genetic × environment interaction in white ash. North Cent. For. Tree. Improv. Conf. 3: 201—208 (1983). —— Kung, F. H. and Clausen, K. E.: Graphic solution in relating seed sources and planting sites for white ash plantations. Silvia Genetica 33: 46—53 (1984). —— Lambeth, C. C.: Juvenile-mature correlations in *Pinaceae* and implications for early selection. For. Sci. 26: 571—580 (1980). —— Neyman, J. and Scott, E. L.: Correction for bias introduced by a transformation of variables. Annals of Math. Stat. 31: 643—655 (1960). —— Rink, G.: Trends in genetic control of juvenile walnut height growth. For. Sci. 30: 821—827 (1984). —— Snyder, E. B.: Glossary for forest tree improvement workers. South. For. Exp. Sta., For. Serv. USDA. 22 p. (1972). —— Wiant, H. V. and Harner, E. J.: Percent bias and standard error in logarithmic regression. For. Sci. 25: 167—168 (1979).