An Analysis of Transformations

By G. E. P. Box and D. R. Cox

University of Wisconsin *Birkbeck College, University of London*

[Read at a Research Methods Meeting of the Society, April 8th, 1964, Professor D. V. Lindley in the Chair]

Summary

In the analysis of data it is often assumed that observations $y_1, y_2, \ldots, y_n$ are independently normally distributed with constant variance and with expectations specified by a model linear in a set of parameters θ. In this paper we make the less restrictive assumption that such a normal, homoscedastic, linear model is appropriate after some suitable transformation has been applied to the y's. Inferences about the transformation and about the parameters of the linear model are made by computing the likelihood function and the relevant posterior distribution. The contributions of normality, homoscedasticity and additivity to the transformation are separated. The relation of the present methods to earlier procedures for finding transformations is discussed. The methods are illustrated with examples.

1. Introduction

The usual techniques for the analysis of linear models as exemplified by the analysis of variance and by multiple regression analysis are usually justified by assuming

(i) simplicity of structure for $E(y)$;
(ii) constancy of error variance;
(iii) normality of distributions;
(iv) independence of observations.

In analysis of variance applications a very important example of (i) is the assumption of additivity, i.e. absence of interaction. For example, in a two-way table it may be possible to represent $E(y)$ by additive constants associated with rows and columns.

If the assumptions (i)–(iii) are not satisfied in terms of the original observations, y, a non-linear transformation of y may improve matters. With this in mind, numerous special transformations for use in the analysis of variance have been examined in the literature; see, in particular, Bartlett (1947). The main emphasis in these studies has tended to be on obtaining a constant error variance, especially when the variance of y is a known function of the mean, as with binomial and Poisson variates.

In multiple regression problems, and in particular in the analysis of response surfaces, assumption (i) might be that $E(y)$ is adequately represented by a rather simple empirical function of the independent variables $x_1, x_2, \ldots, x_l$ and we would want to transform so that this assumption, together with assumptions (ii) and (iii), is approximately satisfied. In some cases transformation of independent as well as of dependent variables might be desirable to produce the simplest possible regression model in the transformed variables. In all cases we are concerned not merely to find a transformation which will justify assumptions but rather to find, where possible, a metric in terms of which the findings may be succinctly expressed.

Each of the considerations (i)–(iii) can, and has been, used separately to select a suitable candidate from a parametric family of transformations. For example, to achieve additivity in the analysis of variance, selection might be based on

(a) minimization of the F value for the degree of freedom for non-additivity (Tukey, 1949); or

(b) minimization of the F ratio for interaction versus error; or

(c) maximization of the F ratio for treatments versus error (Tukey, 1950).

Tukey and Moore (1954) used method (a) in a numerical example, plotting contours of F against (λ_1, λ_2) for transformations in the family $(y+\lambda_2)^{\lambda_1}$. They found that in their particular example the minimizing values were very imprecisely determined.

In both (a) and (b) the general object is to look for a scale on which effects are additive, i.e. to see whether an apparent interaction is removable by a transformation. Of course, only a particular type of interaction is so removable. Whereas (a) can be applied, for example, to a two-way classification without replication, method (b) requires the availability of an error term separated from the interaction term. Thus, if applied to a two-way classification, method (b) could only be used when there was some replication within cells. Finally, method (c) can be used even in a one-way analysis to find the scale on which treatment effects are in some sense most sensitively expressed. In particular, Tukey (1950) suggested multivariate canonical analysis of (y, y^2) to find the linear combination $y+\lambda y^2$ most sensitive to treatment effects. Incidentally, care is necessary in using $y+\lambda y^2$ over the wide ranges commonly encountered with data being considered for transformation, for such a transformation is sensible only so long as the value of λ and the values of y are such that the transformation is monotonic.

For transformation to stabilize variance, the usual method (Bartlett, 1947) is to determine empirically or theoretically the relation between variance and mean. An adequate empirical relation may often be found by plotting log of the within-cell variance against log of the cell mean. Another method would be to choose a transformation, within a restricted family, to minimize some measure of the heterogeneity of variance, such as Bartlett's criterion. We are grateful to a referee for pointing out also the paper of Kleczkowski (1949) in which, in particular, approximate fiducial limits for the parameter λ in the transformation of y to $\log(y+\lambda)$ are obtained. The method is to compute fiducial limits for the parameters in the linear relation observed to hold when the within-cell standard deviation is regressed on the cell mean.

Finally, while there is much work on transforming a single distribution to normality, constructive methods of finding transformations to produce normality in analysis of variance problems do not seem to have been considered.

While Anscombe (1961) and Anscombe and Tukey (1963) have employed the analysis of residuals as a means of detecting departures from the standard assumptions, they have also indicated how transformations might be constructed from certain functions of the residuals.

In regression problems, where both dependent and independent variables can be transformed, there are more possibilities to be considered. Transformation of the independent variables (Box and Tidwell, 1962) can be applied without affecting the constancy of variance and normality of error distributions. An important application is to convert a monotonic non-linear regression relation into a linear one. Obviously it is useless to try to linearize a relation which is not monotonic, but a transformation is sometimes useful in such cases, for example, to make a regression relation more nearly quadratic around its maximum.

2. GENERAL REMARKS ON TRANSFORMATIONS

The main emphasis in this paper is on transformations of the dependent variable. The general idea is to restrict attention to transformations indexed by unknown parameters λ, and then to estimate λ and the other parameters of the model by standard methods of inference. Usually λ will be a one-, or at most two-, dimensional parameter, although there is no restriction in principle. Our procedure then leads to an interesting synthesis of the procedures reviewed in Section 1. It is convenient to make first a few general points about transformations.

First, we can distinguish between analyses in which either (a) the particular transformation, λ, is of direct interest, the detailed study of the factor effects, etc., being of secondary concern; or (b) the main interest is in the factor effects, the choice of λ being only a preliminary step. Type (b) is likely to be much the more common. Nevertheless, (a) can arise, for example, in the analysis of a preliminary set of data. Or, again, we may have two factors, A and B, whose main effects are broadly understood, it being required to study the λ, if any, for which there is no interaction between the factors. Here the primary interest is in λ. In case (b), however, we shall need to fix one, or possibly a small number, of λ's and go ahead with the detailed estimation and interpretation of the factor effects on this particular transformed scale. We shall choose λ partly in the light of the information provided by the data and partly from general considerations of simplicity, ease of interpretation, etc. For instance, it would be quite possible for the formal analysis to show that say $\sqrt{y}$ is the best scale for normality and constancy of variance, but for us to decide that there are compelling arguments of ease of interpretation for working say with $\log y$. The formal analysis will warn us, however, that changes of variance and non-normality may need attention in a refined and efficient analysis of $\log y$. That is, the method developed below for finding a transformation is useful as a guide, but is, of course, not to be followed blindly. In Section 7 we discuss briefly some of the consequences of interpreting factor effects on a scale chosen in the light of the data.

In regression studies, it is sometimes necessary to take an entirely empirical approach to the choice of a relation. In other cases, physical laws, dimensional analysis, etc., may suggest a particular functional form. Thus, in a study of a chemical system one would expect reaction rate to be proportional to some power of the concentration and to the antilog of the reciprocal of absolute temperature. Again, in many fields of technology relationships of the form

$$y \propto x_1^{\beta_1} \ldots x_l^{\beta_l}$$

are very common, suggesting a log transformation of all variables. In such cases the reasonable thing will often be first to apply the transformations suggested by the prior reasoning, and after that consider what further modifications, if any, are needed. Finally, we may know the behaviour of y when the independent variables x_i tend to zero or infinity, and certainly, if we are hopeful that the model might apply over a wide range, we should consider models that are consistent with such limiting properties of the system.

We can distinguish broadly two types of dependent variable, extensive and non-extensive. The former have a relevant property of physical additivity, the latter not. Thus yield of product per batch is extensive. The failure time of a component would be considered extensive if components are replaced on failure, the main thing of interest being the number of components used in a long time. Properties like temperature, viscosity, quality of product, etc., are not extensive. In the absence of

the sort of prior consideration mentioned in the previous paragraph there is no reason to prefer the initial form of a non-extensive variable to any monotonic function of it. Hence, transformations can be applied freely to non-extensive variables. For extensive variables, however, the population mean of y is the parameter determining the long-run behaviour of the system. Thus in the two examples mentioned above, the total yield of product in a long period and the total number of components used in a very long time are determined respectively by the population mean of yield per batch and the mean failure time per component, irrespective of distributional form.

In a narrowly technological sense, therefore, we are interested in the population mean of y, not of some function of y. Hence we either analyse linearly the untransformed data or, if we do apply a transformation in order to make a more efficient and valid analysis, we convert the conclusions back to the original scale. Even in circumstances where, for immediate application, the original scale y is required, it may be better to think in terms of transformed values in which, say, interactions have been removed.

In general, we can regard the usual formal linear models as doing two things:

(a) specifying the questions to be asked, by defining explicitly the parameters which it is the main object of the analysis to estimate;

(b) specifying assumptions under which the above parameters can be simply and effectively estimated.

If there should be conflict between the requirements for (a) and for (b), it is best to pay most attention to (a), since approximate inference about the most meaningful parameters is clearly preferable to formally "exact" inference about parameters whose definition is in some way artificial. Therefore in selecting a transformation we might often give first attention to simplicity of the model structure, for example to additivity in the analysis of variance. This allows simplicity of description and also the main effect of a factor A, measured on a scale for which there appears to be no interaction with a factor B, often has a reasonable possibility of being valid for levels of B outside those of the initial experiment.

3. TRANSFORMATION OF THE DEPENDENT VARIABLE

We work with a parametric family of transformations from y to $y^{(\lambda)}$, the parameter λ, possibly a vector, defining a particular transformation. Two important examples considered here are

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

and

$$y^{(\lambda)} = \begin{cases} \dfrac{(y+\lambda_2)^{\lambda_1}-1}{\lambda_1} & (\lambda_1 \neq 0), \\ \log(y+\lambda_2) & (\lambda_1 = 0). \end{cases} \tag{2}$$

The transformations (1) hold for $y>0$ and (2) for $y>-\lambda_2$. Note that since an analysis of variance is unchanged by a linear transformation (1) is equivalent to

$$y^{(\lambda)} = \begin{cases} y^{\lambda} & (\lambda \neq 0), \\ \log y & (\lambda = 0); \end{cases} \tag{3}$$

the form (1) is slightly preferable for theoretical analysis because it is continuous at $\lambda = 0$. In general, it is assumed that for each λ, $y^{(\lambda)}$ is a monotonic function of y over the admissible range. Suppose that we observe an $n \times 1$ vector of observations $\mathbf{y} = \{y_1, \ldots, y_n\}$, and that the appropriate linear model for the problem is specified by

$$E\{\mathbf{y}^{(\lambda)}\} = \mathbf{a}\boldsymbol{\theta}, \tag{4}$$

where $\mathbf{y}^{(\lambda)}$ is the column vector of *transformed* observations, $\mathbf{a}$ is a known matrix and $\boldsymbol{\theta}$ a vector of unknown parameters associated with the transformed observations.

We now assume that for some unknown λ, the transformed observations $y_i^{(\lambda)}$ $(i = 1, \ldots, n)$ satisfy the full normal theory assumptions, i.e. are independently normally distributed with constant variance σ^2, and with expectations (4). The probability density for the untransformed observations, and hence the likelihood *in relation to these original observations*, is obtained by multiplying the normal density by the Jacobian of the transformation.

The likelihood in relation to the original observations $\mathbf{y}$ is thus

$$\frac{1}{(2\pi)^{\frac{1}{2}n}\sigma^n} \exp\left\{-\frac{(\mathbf{y}^{(\lambda)} - \mathbf{a}\boldsymbol{\theta})'(\mathbf{y}^{(\lambda)} - \mathbf{a}\boldsymbol{\theta})}{2\sigma^2}\right\} J(\lambda; \mathbf{y}), \tag{5}$$

where

$$J(\lambda; \mathbf{y}) = \prod_{i=1}^{n} \left|\frac{dy_i^{(\lambda)}}{dy_i}\right|.$$

We shall examine two ways in which inferences about the parameters in (5) can be made. In the first, we apply "orthodox" large-sample maximum-likelihood theory to (5). This approach leads directly to point estimates of the parameters and to approximate tests and confidence intervals based on the chi-squared distribution.

In the second approach, via Bayes's theorem, we assume that the prior distributions of the θ's and $\log\sigma$ can be taken as essentially uniform over the region in which the likelihood is appreciable and we integrate over the parameters to obtain a posterior distribution for λ; for general discussion of this approach, see, in particular, Jeffreys (1961).

We find the maximum-likelihood estimates in two steps. First, for given λ, (5) is, except for a constant factor, the likelihood for a standard least-squares problem. Hence the maximum-likelihood estimates of the θ's are the least-squares estimates for the dependent variable $y^{(\lambda)}$ and the estimate of σ^2, denoted for fixed λ by $\hat{\sigma}^2(\lambda)$, is

$$\hat{\sigma}^2(\lambda) = \mathbf{y}^{(\lambda)\prime}\mathbf{a}_r\mathbf{y}^{(\lambda)}/n = S(\lambda)/n \tag{6}$$

where, when $\mathbf{a}$ is of full rank,

$$\mathbf{a}_r = \mathbf{I} - \mathbf{a}(\mathbf{a}'\mathbf{a})^{-1}\mathbf{a}', \tag{7}$$

and $S(\lambda)$ is the residual sum of squares in the analysis of variance of $y^{(\lambda)}$.

Thus for fixed λ, the maximized log likelihood is, except for a constant,

$$L_{\max}(\lambda) = -\tfrac{1}{2}n\log\hat{\sigma}^2(\lambda) + \log J(\lambda; \mathbf{y}). \tag{8}$$

In the important special case (1) of the simple power transformation, the second term in (8) is

$$(\lambda - 1)\Sigma\log y_i. \tag{9}$$

In (2), when an unknown origin λ_2 is included, the term becomes

$$(\lambda_1 - 1)\Sigma\log(y_i + \lambda_2). \tag{10}$$

It will now be informative to plot the maximized log likelihood $L_{\max}(\lambda)$ against λ for a trial series of values. From this plot the maximizing value $\hat{\lambda}$ may be read off and we can obtain an approximate $100(1-\alpha)$ per cent confidence region from

$$L_{\max}(\hat{\lambda}) - L_{\max}(\lambda) < \tfrac{1}{2}\chi^2_{\nu_\lambda}(\alpha), \tag{11}$$

where ν_λ is the number of independent components in λ. The main arithmetic consists in doing the analysis of variance of $\mathbf{y}^{(\lambda)}$ for each chosen λ.

If it were ever desired to determine $\hat{\lambda}$ more precisely this could be done by determining numerically the value $\hat{\lambda}$ for which the derivatives with respect to λ are all zero. In the special case of the one parameter power transformation $y^{(\lambda)} = (y^\lambda - 1)/\lambda$,

$$\frac{d}{d\lambda}L_{\max}(\lambda) = -n\frac{\mathbf{y}^{(\lambda)\prime}\mathbf{a}_r\mathbf{u}^{(\lambda)}}{\mathbf{y}^{(\lambda)\prime}\mathbf{a}_r\mathbf{y}^{(\lambda)}} + \frac{n}{\lambda} + \Sigma \log y_i, \tag{12}$$

where $\mathbf{u}^{(\lambda)}$ is the vector of components $\{\lambda^{-1}y_i^\lambda \log y_i\}$. The numerator in (12) is the residual sum of products in the analysis of covariance of $\mathbf{y}^{(\lambda)}$ and $\mathbf{u}^{(\lambda)}$.

The above results can be expressed very simply if we work with the normalized transformation

$$\mathbf{z}^{(\lambda)} = \mathbf{y}^{(\lambda)}/J^{1/n},$$

where $J = J(\lambda; \mathbf{y})$. Then

$$L_{\max}(\lambda) = -\tfrac{1}{2}n \log \hat{\sigma}^2(\lambda; \mathbf{z}),$$

where

$$\hat{\sigma}^2(\lambda; \mathbf{z}) = \frac{\mathbf{z}^{(\lambda)\prime}\mathbf{a}_r\mathbf{z}^{(\lambda)}}{n} = \frac{S(\lambda; \mathbf{z})}{n},$$

where $S(\lambda; \mathbf{z})$ is the residual sum of squares of $\mathbf{z}^{(\lambda)}$. The maximized likelihood is thus proportional to $\{S(\lambda; \mathbf{z})\}^{-n}$ and the maximum-likelihood estimate is obtained by minimizing $S(\lambda; \mathbf{z})$ with respect to λ.

For the simple power transformation

$$z^{(\lambda)} = \frac{y^\lambda - 1}{\lambda \dot{y}^{\lambda-1}},$$

where $\dot{y}$ is the geometric mean of the observations.

For the power transformation with shifted location

$$z^{(\lambda)} = \frac{(y+\lambda_2)^{\lambda_1} - 1}{\lambda_1\{\mathrm{gm}\,(y+\lambda_2)\}^{\lambda_1-1}},$$

where $\mathrm{gm}\,(y+\lambda_2)$ is the sample geometric mean of the $(y+\lambda_2)$'s.

Consider now the corresponding Bayesian analysis. Let the degrees of freedom for residual be $\nu_r = n - \mathrm{rank}\ (\mathbf{a})$, and let

$$s^2(\lambda) = \frac{\mathbf{y}^{(\lambda)\prime}\mathbf{a}_r\mathbf{y}^{(\lambda)}}{\nu_r} = \frac{S(\lambda)}{\nu_r} \tag{13}$$

be the residual mean square in the analysis of variance of $\mathbf{y}^{(\lambda)}$; note the distinction between $\hat{\sigma}^2(\lambda)$, the maximum-likelihood estimate with divisor n, and $s^2(\lambda)$ the "usual"

estimate, with divisor the degrees of freedom ν_r. We first rewrite the likelihood (5), i.e. the conditional probability density function of the y's given $\boldsymbol{\theta}$, σ^2, λ, in the form

$$p(\mathbf{y}\,|\,\boldsymbol{\theta}, \sigma^2, \lambda) = \frac{1}{(2\pi)^{\frac{1}{2}n}\sigma^n} \exp\left\{-\frac{\nu_r s^2(\lambda) + (\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}_\lambda)'\mathbf{a}'\mathbf{a}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}_\lambda)}{2\sigma^2}\right\} J(\lambda; \mathbf{y}), \tag{14}$$

where $\hat{\boldsymbol{\theta}}_\lambda$ is the least-squares estimate of $\boldsymbol{\theta}$ for given λ.

Now consider the choice of the joint prior distribution for the unknown parameters. We first parametrize so that the θ's are linearly independent and hence $n-\nu_r$ in number. Let $p_0(\lambda)$ denote the marginal prior density of λ. We assume that it is reasonable, when making inferences about λ, to take the conditional prior distribution of the θ's and $\log\sigma$, given λ, to be effectively uniform over the range for which the likelihood is appreciable. That is, the conditional prior element given λ is

$$g(\lambda)\, d\boldsymbol{\theta}_\lambda\, d(\log\sigma_\lambda), \tag{15}$$

where, for definiteness, we for the moment denote the effects and variance measured in terms of $y^{(\lambda)}$ by a suffix λ. The factor $g(\lambda)$ is included because the general size and range of the transformed observations $y^{(\lambda)}$ may depend strongly on λ. If the conditional prior distribution (15) were assumed independent of λ, nonsensical results would be obtained.

To determine $g(\lambda)$ we argue as follows. Fix a standard reference value of λ, say λ_1. Suppose provisionally that, for fixed λ, the relation between $y^{(\lambda)}$ and $y^{(\lambda_1)}$ over the range of the observations is effectively linear, say

$$y^{(\lambda)} = \text{const} + l_\lambda y^{(\lambda_1)}. \tag{16}$$

We can then choose $g(\lambda)$ so that when (16) holds, the conditional prior distributions (15) are consistent with one another for different values of λ. In fact, we shall need to apply the answer when the transformations are appreciably non-linear, so that (16) does not hold. There may be a better approach to the choice of a prior distribution than the present one.

It follows from (16) that

$$\log\sigma^2_\lambda = \text{const} + \log\sigma^2_{\lambda_1} \tag{17}$$

and hence, to this order, the prior density of σ^2_λ is independent of λ. However, the θ_λ's are linear combinations of the expected values of the $y^{(\lambda)}$'s, so that

$$\frac{d\theta_\lambda}{d\theta_{\lambda_1}} = l_\lambda.$$

Since there are $n-\nu_r$ independent components to $\boldsymbol{\theta}$, it follows that $g(\lambda)$ is proportional to $1/l_\lambda^{n-\nu_r}$.

Finally we need to choose l_λ. In passing from λ_1 to λ, a small element of volume of the n dimensional sample space is multiplied by $J(\lambda; \mathbf{y})/J(\lambda_1; \mathbf{y})$. An average scale change for a single y component is the nth root of this and, since λ_1 is only a standard reference value, we have approximately

$$l_\lambda = \{J(\lambda; \mathbf{y})\}^{1/n}. \tag{18}$$

Thus, approximately, the conditional prior density (15) is

$$\frac{d\boldsymbol{\theta}_\lambda\, d(\log\sigma_\lambda)}{\{J(\lambda; \mathbf{y})\}^{(n-\nu_r)/n}}.$$

The combined prior element of probability is thus

$$\frac{d\boldsymbol{\theta}\, d(\log\sigma)}{\{J(\lambda;\,\mathbf{y})\}^{(n-\nu_r)/n}}\, p_0(\lambda)\, d\lambda, \tag{19}$$

where we now suppress the suffix λ on $\boldsymbol{\theta}$ and σ.

This is only an approximate result. In particular, the choice of (18) is somewhat arbitrary. However, when a useful amount of information is actually available from the data about the transformation, the likelihood will dominate and the exact choice of (19) is not critical. The prior distribution (19) is interesting in that the observations enter the approximate standardizing coefficient $J(\lambda;\,\mathbf{y})$.

We now have the likelihood (14) and the prior density (19) and can apply Bayes's theorem to obtain the marginal posterior distribution of λ in the form

$$K'_y \frac{L(\lambda|\,y)\, p_0(\lambda)}{\{J(\lambda;\,y)\}^{(n-\nu_r)/n}}, \tag{20}$$

where K'_y is a normalizing constant independent of λ, chosen so that (20) integrates to one with respect to λ, and

$$L(\lambda|\,y) = \int_{-\infty}^{\infty} d(\log\sigma) \int_{-\infty}^{\infty} d\boldsymbol{\theta}\, p(\mathbf{y}|\,\boldsymbol{\theta}, \sigma^2, \lambda). \tag{21}$$

The integral (21) can be evaluated to give

$$L(\lambda|\,y) = \frac{|\mathbf{a}'\mathbf{a}|^{-\frac{1}{2}}\, 2^{\frac{1}{2}\nu_r}\, \Gamma(\frac{1}{2}\nu_r)}{(2\pi)^{\frac{1}{2}\nu_r}\{s^2(\lambda)\}^{\frac{1}{2}\nu_r}\, \nu_r^{\frac{1}{2}\nu_r}} J(\lambda;\,y).$$

Substituting into (20), we have that the posterior distribution of λ is

$$K_y \frac{\{J(\lambda;\,\mathbf{y})\}^{\nu_r/n}}{\{s^2(\lambda)\}^{\frac{1}{2}\nu_r}}\, p_0(\lambda),$$

where K_y is a normalizing constant independent of λ.

Thus the contribution of the observations to the posterior distribution of λ is represented by the factor

$$\{J(\lambda;\,y)\}^{\nu_r/n}/\{s^2(\lambda)\}^{\frac{1}{2}\nu_r}$$

or, on a log scale, by the addition of a term

$$L_b(\lambda) = -\tfrac{1}{2}\nu_r \log s^2(\lambda) + (\nu_r/n)\log J(\lambda;\,y) \tag{22}$$

to $\log p_0(\lambda)$.

Once again if we work with the normalized transformation $z^{(\lambda)} = y^{(\lambda)}/J^{1/n}$, the result is expressed with great simplicity, for

$$L_b(\lambda) = -\tfrac{1}{2}\nu_r \log s^2(\lambda;\,\mathbf{z}) \tag{23}$$

and the posterior density is

$$p(\lambda) = \text{const} \times p_0(\lambda) \times \{S(\lambda;\,\mathbf{z})\}^{-\frac{1}{2}\nu_r}.$$

In practice we can plot $\{S(\lambda; \mathbf{z})\}^{-\frac{1}{2}\nu_r}$ against λ, combining it with any prior information about λ. When the prior density of λ can be taken as locally uniform, the posterior distribution is obtained directly by plotting

$$p_u(\lambda) = k\{S(\lambda; \mathbf{z})\}^{-\frac{1}{2}\nu_r}, \tag{24}$$

where k is chosen to make the total area under the curve unity.

We normally end by selecting a value of λ in the light both of this plot and of other relevant considerations discussed in Section 2. We then proceed to a standard analysis using the indicated transformation.

The maximized log likelihood and the log of the contribution to the posterior distribution of λ may be written respectively as

$$L_{\max}(\lambda) = -\tfrac{1}{2}n\log\{S(\lambda; \mathbf{z})/n\}, \quad L_b(\lambda) = -\tfrac{1}{2}\nu_r\log\{S(\lambda; \mathbf{z})/\nu_r\}.$$

They differ only by substitution of ν_r for n. They are both monotonic functions of $S(\lambda; \mathbf{z})$ and their maxima both occur when the sum of squares $S(\lambda; \mathbf{z})$ is minimized. For general description, $L_{\max}(\lambda)$ and $L_b(\lambda)$ are substantially equivalent. However, it can easily happen that ν_r/n is appreciably less than one, even when n is quite large. Therefore, in applications, the difference cannot always be ignored, especially when a number of models are simultaneously considered.

There are some reasons for thinking $L_b(\lambda)$ preferable to $L_{\max}(\lambda)$ from a non-Bayesian as well as from a Bayesian point of view; see, for example, the introduction by Bartlett (1937) of degrees of freedom into his test for the homogeneity of variance. The general large-sample theorems about the sampling distributions of maximum-likelihood estimates, and the maximum-likelihood ratio chi-squared test, apply just as much to $L_b(\lambda)$ as to $L_{\max}(\lambda)$.

4. TWO EXAMPLES

We have supposed that after suitable transformation from y to $y^{(\lambda)}$, (a) the expected values of the transformed observations are described by a model of simple structure; (b) the error variance is constant; (c) the observations are normally distributed. Then we have shown that the maximized likelihood for λ, and also the approximate contribution to the posterior distribution of λ, are each proportional to a negative power of the residual sum of squares for the variate $z^{(\lambda)} = y^{(\lambda)}/J^{1/n}$.

The "overall" procedure seeks a set of transformation parameters λ for which (a), (b) and (c) are simultaneously satisfied, and sample information on all three aspects goes into the choice. In this Section we now apply this overall procedure to two examples. In Section 5 we shall show how further analysis can show the separate contributions of (a), (b) and (c) in the choice of the transformation. We shall then illustrate this separation using the same two examples.

The above procedure depends on specific assumptions, but it would be quite wrong for fruitful application to regard the assumptions as final. The proper attitude of sceptical optimism is accurately expressed by saying that we tentatively entertain the basis for analysis, rather than that we assume it. The checking of the plausibility of the present procedure will be discussed in Section 5.

A Biological Experiment using a 3 × 4 Factorial Design with Replication

Table 1 gives the survival times of animals in a 3×4 factorial experiment, the factors being (a) three poisons and (b) four treatments. Each combination of the two factors is used for four animals, the allocation to animals being completely randomized.

We consider the application of a simple power transformation $y^{(\lambda)} = (y^\lambda - 1)/\lambda$. Equivalently we shall actually analyse the standardized variate $z^{(\lambda)} = (y^\lambda - 1)/(\lambda \dot{y}^{\lambda-1})$.

TABLE 1

Survival times (unit, 10 hr) of animals in a 3 × 4 factorial experiment

Poison	*Treatment*			
	A	*B*	*C*	*D*
I	0·31	0·82	0·43	0·45
	0·45	1·10	0·45	0·71
	0·46	0·88	0·63	0·66
	0·43	0·72	0·76	0·62
II	0·36	0·92	0·44	0·56
	0·29	0·61	0·35	1·02
	0·40	0·49	0·31	0·71
	0·23	1·24	0·40	0·38
III	0·22	0·30	0·23	0·30
	0·21	0·37	0·25	0·36
	0·18	0·38	0·24	0·31
	0·23	0·29	0·22	0·33

We are tentatively entertaining the model that after such transformation

(a) the expected value of the transformed variate in any cell can be represented by additive row and column constants, i.e. that no interaction terms are needed,
(b) the error variance is constant,
(c) the observations are normally distributed.

The maximized likelihood and the posterior distribution are functions of the residual sum of squares for $\mathbf{z}^{(\lambda)}$ after eliminating row and column effects. This sum of squares is denoted $S(\lambda; \mathbf{z})$. It has 42 degrees of freedom and is the result of pooling the "within groups" and the "interaction" sums of squares.

Table 2 gives $S(\lambda; \mathbf{z})$ together with $L_{\max}(\lambda)$ and $p_u(\lambda)$ over the interesting ranges. The constant k in $k e^{L_b(\lambda)} = p_u(\lambda)$ is the reciprocal of the area under the curve $Y = e^{L_b(\lambda)}$ determined by numerical integration. Graphs of $L_{\max}(\lambda)$ and of $p_u(\lambda)$ are shown in Fig. 1. This analysis points to an optimal value of about $\hat{\lambda} = -0{\cdot}75$. Using (11) the curve of maximized likelihood gives an approximate 95 per cent confidence interval for λ extending from about $-1{\cdot}13$ to $-0{\cdot}37$.

The posterior distribution $p_u(\lambda)$ is approximately normal with mean $-0{\cdot}75$ and standard deviation 0·22. About 95 per cent of this posterior distribution is included within the limits $-1{\cdot}18$ and $-0{\cdot}32$.

The reciprocal transformation has a natural appeal for the analysis of survival times since it is open to the simple interpretation that it is the *rate of dying* which is to be considered. Our analysis shows that it would, in fact, embody most of the advantages obtainable. The complete analysis of variance for the untransformed data and for the reciprocal transformation (taken in the z form) is shown in Table 3.

Whereas no great change occurs on transformation in the mean squares associated with poisons and treatments, the within groups mean square has shrunk to a third of

TABLE 2

Biological data. Calculations based on an additive, homoscedastic, normal model in the transformed observations

λ	$S(\lambda; \mathbf{z})$	$L_{\max}(\lambda)$	λ	$S(\lambda; \mathbf{z})$	$L_{\max}(\lambda)$
1·0	1·0509	91·72	−1·0	0·3331	119·29
0·5	0·6345	103·83	−1·2	0·3586	117·52
0·0	0·4239	113·51	−1·4	0·4007	114·86
−0·2	0·3752	116·44	−1·6	0·4625	111·43
−0·4	0·3431	118·58	−2·0	0·6639	102·74
−0·6	0·3258	119·82	−2·5	1·1331	89·91
−0·8	0·3225	120·07	−3·0	2·0489	75·69

λ	$p_u(\lambda)$	λ	$p_u(\lambda)$
0·0	0·01	−0·8	1·82
−0·1	0·02	−0·9	1·42
−0·2	0·08	−1·0	0·92
−0·3	0·26	−1·1	0·47
−0·4	0·49	−1·2	0·19
−0·5	0·94	−1·3	0·07
−0·6	1·46	−1·5	0·01
−0·7	1·82		

$L_{\max}(\lambda) = -24 \log \hat{\sigma}^2(\lambda; \mathbf{z}) = \log \{S(\lambda; \mathbf{z})\}^{-24} + 92{\cdot}91$; $p_u(\lambda) = k\, e^{L_b(\lambda)} = 0{\cdot}866 \times 10^{-10}\{S(\lambda; \mathbf{z})\}^{-21}$.

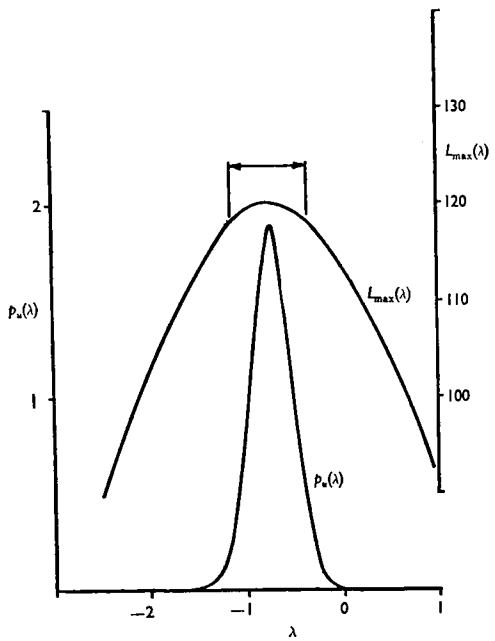

FIG. 1. Biological data. Functions $L_{\max}(\lambda)$ and $p_u(\lambda)$. Arrows show approximate 95 per cent. confidence interval for λ.

its value and the interaction mean square is now much closer in size to that within groups. Thus, in the transformed metric, not only is greater simplicity of interpretation possible but also the sensitivity of the experiment, as measured by the ratios

TABLE 3

Analyses of variance of biological data

	Mean squares × 1000		
	Degrees of freedom	*Untransformed*	*Reciprocal transformation (z form)*
Poisons . .	2	516·5	568·7
Treatments . .	3	307·1	221·9
$P \times T$ . . .	6	41·7	8·5
Within groups .	36	22·2	7·8

of the poisons and the treatments mean squares to the residual square, has been increased almost threefold. We shall not here consider the detailed interpretation of the factor effects.

A Textile Experiment using a Single Replicate of a 3^3 Design

In an unpublished report to the Technical Committee, International Wool Textile Organization, Drs A. Barella and A. Sust described some experiments on the behaviour of worsted yarn under cycles of repeated loading. Table 4 gives the numbers of cycles to failure, y, obtained in a single replicate of a 3^3 experiment in which the factors are

x_1: length of test specimen (250, 300, 350 mm.),
x_2: amplitude of loading cycle (8, 9, 10 mm.),
x_3: load (40, 45, 50 gm.).

In Table 4 the levels of the x's are denoted conventionally by -1, 0, 1.

It is useful to describe first the results of a rather informal analysis of Table 4. Barella and Sust fitted a full equation of second degree in x_1, x_2 and x_3, but the conclusions were very complicated and messy. In view of the wide relative range of variation of y, it is natural to try analysing instead log y, and there results a great simplification. All linear regression terms are very highly significant and all second-degree terms are small. Further, it is natural to take logs also for the independent variables, i.e. to think in terms of relationships like

$$y \propto x_1^{\beta_1} x_2^{\beta_2} x_3^{\beta_3}. \tag{25}$$

The estimates of the β's, from the linear regression coefficients of log y on the log x's, are, with their estimated standard errors,

$$\hat{\beta}_1 = 4{\cdot}96 \pm 0{\cdot}20, \quad \hat{\beta}_2 = -5{\cdot}27 \pm 0{\cdot}30, \quad \hat{\beta}_3 = -3{\cdot}15 \pm 0{\cdot}30.$$

Since $\hat{\beta}_1 \simeq -\hat{\beta}_2$, the combination $\log x_1 - \log x_2 = \log(x_1/x_2)$ is suggested by the data as of possible importance. In fact, x_2/x_1 is just the fractional amplitude of the loading cycle; indeed, naïve dimensional considerations suggest this as a possible factor, although there are in fact other relevant lengths, so that dependence on x_1

and x_2 separately is not inconsistent with dimensional considerations. If, however, we write $x_2/x_1 = x_4$ and round the regression coefficients, we have the simple formula

$$y \propto x_4^{-5} x_3^{-3}$$

which fits the data remarkably well.

TABLE 4

Cycles to failure of worsted yarn: 3^3 factorial experiment

Factor levels			*Cycles to failure, y*
x_1	x_2	x_3	
−1	−1	−1	674
−1	−1	0	370
−1	−1	+1	292
−1	0	−1	338
−1	0	0	266
−1	0	+1	210
−1	+1	−1	170
−1	+1	0	118
−1	+1	+1	90
0	−1	−1	1,414
0	−1	0	1,198
0	−1	+1	634
0	0	−1	1,022
0	0	0	620
0	0	+1	438
0	+1	−1	442
0	+1	0	332
0	+1	+1	220
+1	−1	−1	3,636
+1	−1	0	3,184
+1	−1	+1	2,000
+1	0	−1	1,568
+1	0	0	1,070
+1	0	+1	566
+1	+1	−1	1,140
+1	+1	0	884
+1	+1	+1	360

In this case, there seem strong general arguments for starting with a log transformation of all variables. Power laws are frequently effective in the physical sciences; also, provided that the signs of the β's are right, (25) has sensible limiting behaviour for $x_2, x_3 \to 0, \infty$; finally, the obvious normal theory model based on transforming (25) gives distributions over positive values of y only.

Nevertheless, it is interesting to see whether the method of the present paper applied directly to the data of Table 4 produces the log transformation. In this paper, transformations of the dependent variable alone are considered; in fact, since the relative range of the x's is not very great, transformation of the x's does not have a big effect on the linearity of the regression.

We first consider the application of a simple power transformation in terms, as before, of the standardized variate $z^{(\lambda)} = (y^{\lambda}-1)/(\lambda\dot{y}^{\lambda-1})$. We tentatively suppose that after such transformation

(a) the expected value of the transformed response can be represented merely by a model *linear* in the x's,
(b) the error variance is constant,
(c) the observations are normally distributed.

The maximized likelihood and the posterior distribution are functions of the residual sum of squares for $z^{(\lambda)}$ after fitting only a linear model to the x's. Since there are four constants in the linear regression model this residual sum of squares has $27-4 = 23$ degrees of freedom; we denote it by $S(\lambda; \mathbf{z})$.

Table 5 shows $S(\lambda; \mathbf{z})$ together with $L_{\max}(\lambda)$ and $p_u(\lambda)$ over the interesting ranges and the results are plotted in Fig. 2. The optimal value for the transformation parameter is $\hat{\lambda} = -0{\cdot}06$. The transformation is determined remarkably closely in this

TABLE 5

Textile data. Calculations based on normal linear model in the transformed observations

λ	$S(\lambda; \mathbf{z})$	$L_{\max}(\lambda)$	λ	$S(\lambda; \mathbf{z})$	$L_{\max}(\lambda)$
1·00	5·4810	21·52	−0·20	0·2920	61·11
0·80	2·9978	29·67	−0·40	0·5478	52·61
0·60	1·5968	38·17	−0·60	1·1035	43·16
0·40	0·8178	47·21	−0·80	2·1396	34·22
0·20	0·4115	56·48	−1·00	3·9955	25·79
0·00	0·2519	63·10			

λ	$p_u(\lambda)$	λ	$p_u(\lambda)$
0·20	0·02	−0·10	4·66
0·15	0·09	−0·15	2·36
0·10	0·42	−0·20	0·77
0·05	1·58	−0·25	0·19
0·00	4·18	−0·30	0·04
−0·05	5·64	−0·35	0·01

$L_{\max}(\lambda) = -13{\cdot}5 \log \hat{\sigma}^2(\lambda; \mathbf{z}) = \{S(\lambda; \mathbf{z})\}^{-13{\cdot}5} + 44{\cdot}49.$
$p_u(\lambda) = k\, e^{L_b(\lambda)} = 0{\cdot}540 \times 10^{-6}\{S(\lambda; \mathbf{z})\}^{-11{\cdot}5}.$

example, the approximate 95 per cent confidence range extending only from −0·18 to +0·06. The posterior distribution $p_u(\lambda)$ has its mean at −0·06. About 95 per cent of the distribution is included between −0·20 and +0·08. As we have mentioned, the advantages of a log transformation corresponding to the choice $\lambda = 0$ are very great and such a choice is now seen to be strongly supported by the data.

The complete analysis of variance for the untransformed and the log transformation, taken in the z form, is shown in Table 6.

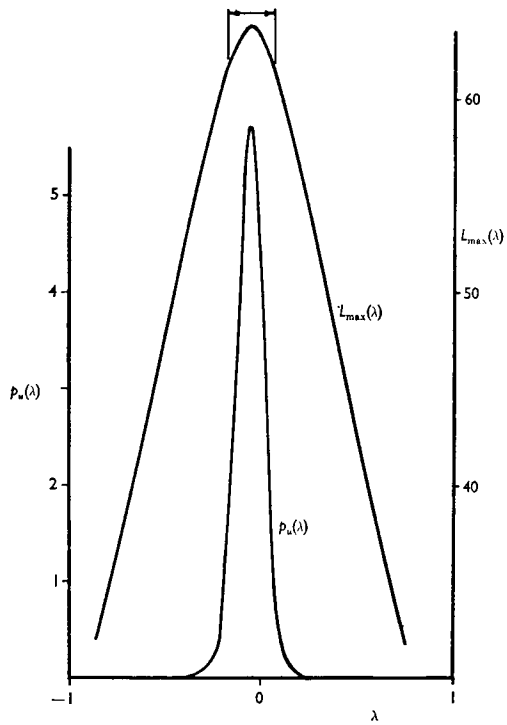

FIG. 2. Textile data. Functions $L_{\max}(\lambda)$ and $p_u(\lambda)$. Arrows show approximate 95 per cent confidence interval for λ.

TABLE 6

Analyses of variance of textile data

		Mean squares × 1000	
	Degrees of freedom	*Untransformed*	*Logarithmic transformation (z form)*
Linear . .	3	4,916·2	2,374·4
Quadratic . .	6	704·1	8·1
Residual . .	17	73·9	11·9

The transformation eliminates the need for second-order terms in the regression equation while at the same time increasing the sensitivity of the analysis by about three, as judged by the ratio of linear and residual mean squares.

For this example we have also tried out the procedures we have discussed using the two parameter transformation $y^{(\lambda)} = \{(y+\lambda_2)^{\lambda_1} - 1\}/\lambda_1$ or in the z form actually

used here $z^{(\lambda)} = \{(y+\lambda_2)^{\lambda_1} - 1\}/\{\lambda_1 \text{gm}(y+\lambda_2)\}^{\lambda_1 - 1}$. Incidentally the calculation and print out of 77 analysis of variance tables, involving in each case the fitting of a general equation of second degree, and calculation of residuals and fitted values took 2 min. 6 sec. on the C.D.C. 1604 electronic computer. The full numerical results can be obtained from the authors, but are not given here. Instead approximate contours of $-11{\cdot}5 \log S(\lambda; \mathbf{z})$, and hence of $S(\lambda; \mathbf{z})$ itself, of the maximized likelihood and of $p_u(\lambda_1, \lambda_2)$, are shown in Fig. 3. If the joint posterior distribution $p_u(\lambda_1, \lambda_2)$ were normal then a region which excluded 100α per cent of the total posterior probability could be given by

$$L_b(\hat{\lambda}_1, \hat{\lambda}_2) - L_b(\lambda_1, \lambda_2) = \chi^2_2(\alpha). \tag{26}$$

The shape of the contours indicates that the normal assumption is not very exact. Nevertheless, the quantity 100α obtained from (26) has been used to label the contours in Fig. 3 which thus roughly indicates the posterior probability distribution. For this example no appreciable improvement results from the addition of the further transformation parameter λ_2.

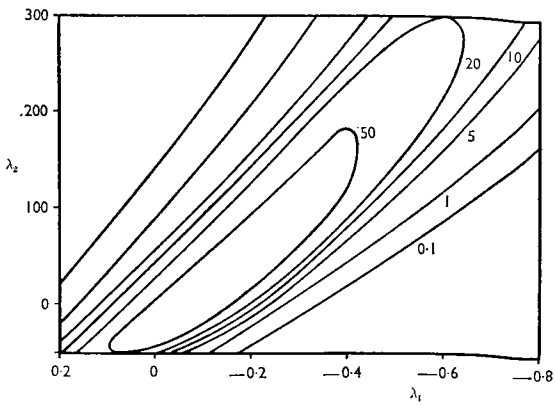

FIG. 3. Textile data. Transformation to $(y+\lambda_2)^{\lambda_1}$. Contours of $p_u(\lambda_1, \lambda_2)$ labelled with approximate percentage of posterior distribution excluded.

5. FURTHER ANALYSIS OF THE TRANSFORMATION

5.1. *General Procedure for Further Analysis*

The general procedure discussed above seeks to achieve simultaneously a model with (a) simple structure for the expectations, (b) constant variance and (c) normal distributions. Further analysis is sometimes profitable to see the separate contributions of these three elements to the transformation. Such analysis may indicate

(i) how simple a model we are justified in using;
(ii) what weight is given to the considerations (a) – (c) in choosing λ;
(iii) whether different transformations are really needed to achieve the different objectives and hence whether or not the value of λ chosen using the overall procedure is a compatible compromise.

Of course, quite often careful inspection of the data will answer (i)–(iii) adequately for practical purposes. Nevertheless, a further analysis is of interest.

We aim at simplicity both to achieve ease of understanding and to allow an efficient analysis. Validity of the formal tests associated with analysis of variance may, in virtue of the robustness of these tests, often hold to a good enough approximation even with the untransformed data. We stress, however, that such approximate validity is not by itself enough to justify an analysis; sensitivity must be considered as well as robustness. Thus in the biological example we have about one-third the sensitivity on the original scale as on the transformed scale. The approximate validity of significance tests on the original scale would be very poor consolation for the substantial loss of information involved in using the untransformed analysis. In any case even such validity is usually only preserved under the null hypothesis that all treatment effects are zero.

For the further analysis we again explore two approaches, one via maximum likelihood and the other via Bayes's theorem. Consider a general model to which a constraint C can be applied or relaxed, so that the relative merits of the simple and of the more complex model can be assessed. For example, the general model may include interaction terms, the constraint C being that the interaction terms are zero.

If $L_{\max}(\lambda)$ and $L_{\max}(\lambda|C)$ denote maximized log likelihoods for the general model and for the constrained model, then

$$L_{\max}(\lambda|C) = L_{\max}(\lambda) + \{L_{\max}(\lambda|C) - L_{\max}(\lambda)\}. \tag{27}$$

Here the second term on the right-hand side is a statistic for testing for the presence of the constraint.

More generally, with a succession of constraints, we have

$$\begin{aligned} L_{\max}(\lambda|C_1, C_2) = L_{\max}(\lambda) &+ \{L_{\max}(\lambda|C_1) - L_{\max}(\lambda)\} \\ &+ \{L_{\max}(\lambda|C_1, C_2) - L_{\max}(\lambda|C_1)\}, \end{aligned} \tag{28}$$

and the three terms on the right of (28) can be examined separately. The detailed procedure should be clear from the examples to follow.

To apply the Bayesian approach, we write the posterior density of λ

$$p(\lambda|C) = p(\lambda) \times \frac{p(C|\lambda)}{p(C)}, \tag{29}$$

where $p(C) = E_\lambda\{p(C|\lambda)\}$ is a constant independent of λ. That is, the posterior density of λ under the constrained model is the posterior density under the general model multiplied by a factor proportional to the conditional probability of the constraint given λ. Successive factorization can be applied when there is a series of successively applied constraints, giving, for example,

$$p(\lambda|C_1, C_2) = p(\lambda) \times \frac{p(C_1|\lambda)}{p(C_1)} \times \frac{p(C_2|\lambda, C_1)}{p(C_2|C_1)}, \tag{30}$$

where $p(C_2|C_1) = E_\lambda\{p(C_2|\lambda, C_1)\}$ is a further constant independent of λ. Note that we are concerned here not with the probabilities that the constraints are true, but with the contributions of the constraints to the final function $p(\lambda|C_1, C_2)$.

5.2. *Structure of the Expectation*

Now very often the most important question is: how simple a form can we use for $E\{y^{(\lambda)}\}$? Thus in the analysis of the biological example in Section 4, we assumed, among other things, that additivity can be achieved by transformation. In fact,

interaction terms may or may not be needed. Similarly, in our analysis of the textile example we took a linear model with four parameters; the full second-degree model with ten parameters may or may not be necessary.

Now let A, H and N denote respectively the constraints to the simpler linear model (without interaction or second-degree terms), to a heteroscedastic model and to a model with normal distributions. Then,

$$L_{\max}(\lambda|A,H,N) = L_{\max}(\lambda|H,N)+\{L_{\max}(\lambda|A,H,N)-L_{\max}(\lambda|H,N)\}. \tag{31}$$

Let the parameter $\boldsymbol{\theta}$ in the expectation under the general linear model be partitioned $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ where $\boldsymbol{\theta}_2 = 0$ is the constraint A. Denote the degrees of freedom associated with $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ by ν_1 and ν_2. If ν_r is the number of degrees of freedom for residual in the complex model, the number in the simpler model is thus $\nu_r+\nu_2$.

As before, we work with the standardized variable $z^{(\lambda)} = y^{(\lambda)}/J^{1/n}$. If we identify residual sums of squares by their degrees of freedom, we have

$$L_{\max}(\lambda|\boldsymbol{\theta}_2 = 0, H, N) = -\tfrac{1}{2}n\log\{S_{\nu_r+\nu_2}(\lambda;\mathbf{z})/n\}, \tag{32}$$

whereas

$$L_{\max}(\lambda|H,N) = -\tfrac{1}{2}n\log\{S_{\nu_r}(\lambda;\mathbf{z})/n\}. \tag{33}$$

Thus, in the textile example, S_{ν_r} refers to the residual sum of squares from a second-degree model and $S_{\nu_r+\nu_2}$ refers to the residual sum of squares from a first-degree model. Quite generally

$$S_{\nu_r+\nu_2}(\lambda;\mathbf{z}) = S_{\nu_r}(\lambda;\mathbf{z})+S_{\nu_2.\nu_1}(\lambda;\mathbf{z}),$$

where $S_{\nu_2.\nu_1}(\lambda;\mathbf{z})$ denotes the extra sum of squares of $\mathbf{z}^{(\lambda)}$ for fitting $\boldsymbol{\theta}_2$, adjusting for $\boldsymbol{\theta}_1$, and has ν_2 degrees of freedom.

Thus with (32) and (33) the decomposition (31) becomes

$$L_{\max}(\lambda|\boldsymbol{\theta}_2 = 0, H, N) = L_{\max}(\lambda|H,N)-\tfrac{1}{2}n\log\left\{1+\frac{\nu_2}{\nu_r}F(\lambda;\mathbf{z})\right\}, \tag{34}$$

where

$$F(\lambda;\mathbf{z}) = \frac{S_{\nu_2.\nu_1}(\lambda;\mathbf{z})/\nu_2}{S_{\nu_r}(\lambda;\mathbf{z})/\nu_r} \tag{35}$$

is the standard F ratio, in the analysis of variance of $\mathbf{z}^{(\lambda)}$, for testing the restriction to the simpler model.

Equation (34) thus provides an analysis of the overall criterion into a part taking account only of homoscedasticity (H) and normality (N) plus a part representing the additional requirement of a simple linear model, given that H and N have been achieved.

In the corresponding Bayesian analysis (30) gives

$$p(\lambda|\boldsymbol{\theta}_2 = 0, H, N) = p(\lambda|H,N)\times k_A\, p(\boldsymbol{\theta}_2 = 0|\lambda, H, N), \tag{36}$$

where

$$1/k_A = E_{\lambda|H,N}\{p(\boldsymbol{\theta}_2 = 0|\lambda, H, N)\},$$

the expectation being taken over the distribution $p(\lambda|H,N)$.

Note that since the condition $\boldsymbol{\theta}_2 = 0$ is given, there is no component for these parameters in the prior distribution, so that the left-hand side of (36) is the posterior density obtained previously assuming A. Thus, in terms of the standardized variable $\mathbf{z}^{(\lambda)}$, the left-hand side is

$$p_0(\lambda)\, C_{\nu_r+\nu_2}\{S_{\nu_r+\nu_2}(\lambda;\mathbf{z})\}^{-\frac{1}{2}(\nu_r+\nu_2)}, \tag{37}$$

where the normalizing constant is given by

$$C^{-1}_{\nu_r+\nu_2} = \int p_0(\lambda)\{S_{\nu_r+\nu_2}(\lambda;\,\mathbf{z})\}^{-\frac{1}{2}(\nu_r+\nu_2)}\,d\lambda.$$

Similarly, in the general model with $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ both free to vary, we obtain the first factor on the right-hand side of (36) as

$$p(\lambda\,|\,H, N) = p_0(\lambda)\,C_{\nu_r}\{S_{\nu_r}(\lambda;\,\mathbf{z})\}^{-\frac{1}{2}\nu_r}, \tag{38}$$

with

$$C^{-1}_{\nu_r} = \int p_0(\lambda)\{S_{\nu_r}(\lambda;\,\mathbf{z})\}^{-\frac{1}{2}\nu_r}\,d\lambda.$$

Thus, from (37) and (38), the second factor on the right-hand side of (36) must be

$$\frac{C_{\nu_r+\nu_2}}{C_{\nu_r}}\frac{\{S_{\nu_r}(\lambda;\,\mathbf{z})\}^{\frac{1}{2}\nu_r}}{\{S_{\nu_r+\nu_2}(\lambda;\,\mathbf{z})\}^{\frac{1}{2}(\nu_r+\nu_2)}}. \tag{39}$$

Now the general equation (36) shows that this last expression must be proportional to $p(\boldsymbol{\theta}_2 = 0\,|\,\lambda, H, N)$. It is worth proving this directly. To do this, consider a transformed scale on which constant variance and normality have been attained and the standard estimates $\hat{\boldsymbol{\theta}}_2$ and s^2 calculated. For the moment, we need not indicate explicitly the dependence on λ and z. We denote the matrix of the reduced least-squares equations for $\boldsymbol{\theta}_2$, eliminating $\boldsymbol{\theta}_1$, by $\mathbf{b}$, so that the covariance matrix of $\boldsymbol{\theta}_2$ is $\sigma^2\mathbf{b}^{-1}$. The elements of $\mathbf{b}$ and $\mathbf{b}^{-1}$ are denoted b_{ij} and b^{ij}. Also we write $\rho_{ij} = b^{ij}/\sqrt{(b^{ii}b^{jj})}$ and $\{\rho^{ij}\}$ for the matrix inverse to $\{\rho_{ij}\}$. Then the joint distribution of

$$t_i = \frac{\theta_{2i} - \hat{\theta}_{2i}}{s\sqrt{b^{ii}}}$$

is (Cornish, 1954; Dunnett and Sobel, 1954)

$$\text{const} \times \left(1 + \frac{\Sigma\rho^{ij}t_i t_j}{\nu_r}\right)^{-\frac{1}{2}(\nu_r+\nu_2)}$$

where here and later the constant involves neither the parameters nor the observations. With uniform prior distributions for the θ's and for log σ, this is also the posterior distribution of the quantities $(\theta_{2i} - \hat{\theta}_{2i})/(s/\sqrt{b^{ii}})$, where now the θ_{2i} are the random variables. Transforming from the t_i's to the θ_{2i}'s we have that

$$p(\boldsymbol{\theta}_2\,|\,\lambda, H, N) = \text{const} \times (s^2_{\nu_r})^{-\frac{1}{2}\nu_2}\left\{1 + \frac{(\boldsymbol{\theta}_2 - \hat{\boldsymbol{\theta}}_2)'\mathbf{b}(\boldsymbol{\theta}_2 - \hat{\boldsymbol{\theta}}_2)}{\nu_r s^2_{\nu_r}}\right\}^{-\frac{1}{2}(\nu_r+\nu_2)}$$

whence

$$p(\boldsymbol{\theta}_2 = 0\,|\,\lambda, H, N) = \text{const} \times (S_{\nu_r})^{-\frac{1}{2}\nu_2}\left\{1 + \frac{\hat{\boldsymbol{\theta}}_2'\mathbf{b}\,\hat{\boldsymbol{\theta}}_2}{S_{\nu_r}}\right\}^{-\frac{1}{2}(\nu_r+\nu_2)}$$

$$= \text{const} \times \frac{S^{\frac{1}{2}\nu_r}_{\nu_r}}{(S_{\nu_r+\nu_2})^{\frac{1}{2}(\nu_r+\nu_2)}}. \tag{40}$$

If now we restore in our notation the dependence on λ, comparison of (40) with (39) proves the required result; the appropriateness of the constant is easily checked.

Thus (36) provides an analysis of the overall density into a part $p(\lambda\,|\,H, N)$ taking account only of homoscedasticity and normality, and a second part, (39), in which the influence of the simplifying constraint is measured.

Equation (39) can be rewritten

$$\text{const} \times \{S_{\nu_r}(\lambda; \mathbf{z})\}^{-\frac{1}{2}\nu_2} \left\{1 + \frac{\nu_2}{\nu_r} F(\lambda; \mathbf{z})\right\}^{-\frac{1}{2}(\nu_r+\nu_2)}. \tag{41}$$

Now, by (34), the corresponding expression in the maximum-likelihood approach is given, in a logarithmic version, by

$$-\tfrac{1}{2}n \log \left\{1 + \frac{\nu_2}{\nu_r} F(\lambda; \mathbf{z})\right\}. \tag{42}$$

The essential difference between (41) and (42) is the occurrence of the term in $S_{\nu_r}(\lambda; \mathbf{z})$ in (41). In conventional large sample theory, ν_r is supposed large compared with ν_2 and then in the limit the variation with λ of the additional term is negligible, and the effect of both terms can be represented by plotting the standard F ratio as a function of λ. In applications, however, ν_2/ν_r may well be appreciable; thus in the textile example $\nu_2/\nu_r = 6/17$.

Hence (41) and (42) could lead to appreciably different conclusions, for example, if we found a particular value of λ giving a low value of $F(\lambda; z)$ but a relatively high value of $S_{\nu_r}(\lambda; \mathbf{z})$.

The distinction between (41) and (42) from a Bayesian point of view can be expressed as follows. In (41) there occurs the *ordinate* of the posterior distribution of $\boldsymbol{\theta}_2$ at $\boldsymbol{\theta}_2 = 0$. On the other hand, the F ratio, which determines (42), is a monotonic function of the *probability mass* outside the contour of the posterior distribution passing through $\boldsymbol{\theta}_2 = 0$. Alternatively, a calculation of the posterior probability of a small region near $\boldsymbol{\theta}_2 = 0$ having a length proportional to σ_z in each of the ν_2 component directions gives an expression equivalent to (42). The difference between (41) and (42) will be most pronounced if there exists an extreme transformation producing a low value of $F(\lambda; z)$ but a large value of $S_{\nu_r}(\lambda; z)$, corresponding to a large spread of the posterior distribution of $\boldsymbol{\theta}_2$. Expression (42) would give an answer tending to favour this transformation, whereas (41) would not.

5.3. *Application to Textile Example*

We now illustrate the above analysis using the textile data. The calculations are set out in Table 7 and displayed in Figs. 4 and 5. We discuss the conclusions in some detail here. In practice, however, the most useful aspect of this approach is the opportunity for graphical assessment.

Fig. 4 shows that the curvature of $L_{\max}(\lambda | H, N)$ is much less than that of $L_{\max}(\lambda | A, H, N)$ previously given in Fig. 2, the constraint A here being that the second-degree terms are supposed zero. The inequality

$$L_{\max}(\hat{\lambda} | H, N) - L_{\max}(\lambda | H, N) < \tfrac{1}{2}\chi_1^2(\alpha) \tag{43}$$

thus gives the much wider approximate 95 per cent confidence interval $(-0{\cdot}48, 0{\cdot}13)$ for λ indicated by HN in Fig. 4 and compared with the previous interval, marked AHN. Since the constraint has 6 degrees of freedom the sampling distribution of

$$-2\{L_{\max}(\lambda | A, H, N) - L_{\max}(\lambda | H, N)\} \tag{44}$$

for fixed normalizing λ is asymptotically χ_6^2. Alternatively, (44), being a monotonic function of F, can be tested exactly. Thus we can decide for which λ's, if any, the inclusion of the constraint is compatible with the data. In Fig. 5, $F(\lambda; \mathbf{z})$ is close to

unity over the interesting range of λ close to zero, so that we can use the simpler model in this neighbourhood. The range indicated by C in Fig. 4 is that for which F is less than 2·70, the 5 per cent significance point.

TABLE 7

Textile data. Calculations for the analysis of the transformation

λ	$L_{\max}(\lambda \mid A, H, N)$	$L_{\max}(\lambda \mid H, N)$	*Difference* $= -13{\cdot}5\times$	
			$\log(1+\frac{6}{7}F(\lambda; z))$	$F(\lambda; z)$
1·00	21·52	41·41	−19·89	9·52
0·80	29·67	49·14	−19·47	9·15
0·60	38·17	55·65	−17·48	7·50
0·40	47·21	60·59	−13·38	4·80
0·20	56·48	63·99	− 7·51	2·09
0·00	63·10	66·02	− 2·92	0·68
−0·20	61·11	66·89	− 5·78	1·51
−0·40	52·61	66·07	−13·46	4·84
−0·60	43·16	62·68	−19·52	9·19
−0·80	34·22	56·44	−22·22	11·85
−1·00	25·79	48·18	−22·39	·12·03

λ	$p_u(\lambda \mid A, H, N)$	$p_u(\lambda \mid H, N)$	$k_A\, p_u(A \mid \lambda, H, N)$
0·20	0·02	0·32	0·05
0·15	0·09	0·49	0·18
0·10	0·42	0·69	0·62
0·05	1·58	0·93	1·71
0·00	4·18	1·19	3·51
−0·05	5·64	1·47	3·84
−0·10	4·66	1·76	2·65
−0·15	2·36	1·96	1·20
−0·20	0·77	2·06	0·37
−0·25	0·19	2·03	0·09
−0·30	0·04	1·88	0·02
−0·35	0·01	1·59	0·01

The Bayesian analysis follows parallel lines. In Fig. 4, $p_u(\lambda \mid H, N)$ has a much greater spread than $p_u(\lambda \mid A, H, N)$. Fig. 5 shows $p_u(\lambda \mid H, N)$ with the component $k_A p(A \mid \lambda, H, N)$ from the constraint. When multiplied together they give the overall density $p_u(\lambda \mid A, H, N)$. A value of λ near zero maximizes the posterior density assuming the constraint and is consistent with the information in $p_u(\lambda \mid H, N)$.

There is, however, nothing in our Bayesian analysis itself to tell us whether the simplified model with the constraint is compatible with the data, even for the best possible λ. There is an important general point here. All probability calculations in statistical inference are conditional in one way or another. In particular, Bayesian posterior distributions such as $p_u(\lambda \mid A, H, N)$ are conditional on the model, in particular here on assumption A. It could easily happen that there is no value of λ for which A is at all reasonable, but to check on this we need to supplement the

Bayesian argument (Anscombe, 1961). Here we can do this by a significance test based on the sampling distribution of a suitable function of the observations, namely $F(\lambda; \mathbf{z})$. For λ around zero the value of $F(\lambda; z)$ is, in fact, well within the significance limits, so that we can reasonably use the posterior distribution of λ in question.

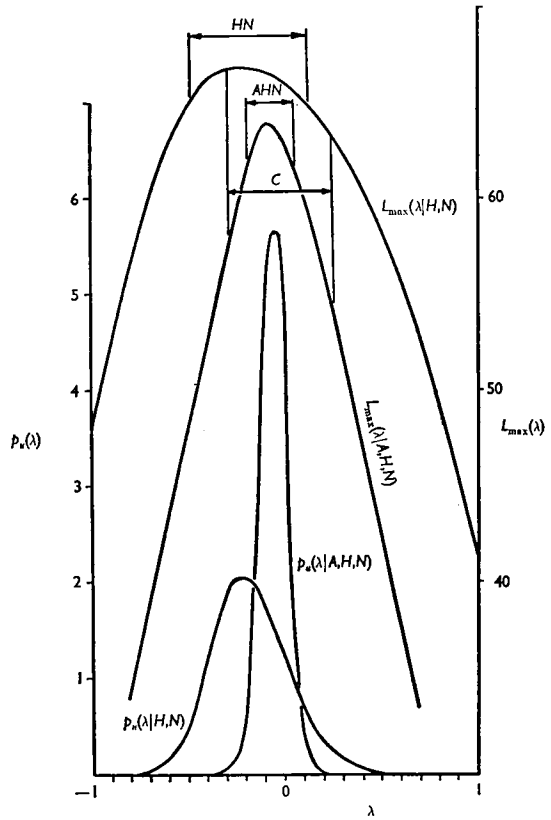

FIG. 4. Textile data. Functions $L_{max}(\lambda)$ and $p_u(\lambda)$ under different models. *A*: additivity. *H*: homogeneity of variance. *N*: normality. Arrows *HN*, *AHN* show approximate 95 per cent confidence intervals for λ. Arrows C show range for which *F* for second-degree terms is not significant at 5 per cent level.

5.4. *Homogeneity of Variance*

Suppose that we have k groups of data, the expectation and variance being constant within each group. In the lth group, let the variance be σ_l^2 and let $S^{(l)}$ denote the sum of squares of deviations, having $\nu_l = n_l - 1$ degrees of freedom. Write $\Sigma n_l = n$, $\Sigma \nu_l = n-k$. Thus in our biological example, $k = 12$, $\nu_1 = \ldots = \nu_{12} = 3$, $n_1 = \ldots = n_{12} = 4$ and $\nu = 36$, $n = 48$.

Now suppose that a transformation to $y^{(\lambda)}$ exists which induces normality simultaneously in all groups. Then in terms of the standardized variable $z^{(\lambda)}$, the maximized log likelihood is

$$L_{max}(\lambda|N) = -\tfrac{1}{2}\Sigma n_l \log\{S^{(l)}(\lambda; \mathbf{z})/n_l\}, \tag{45}$$

where $S^{(l)}(\lambda; \mathbf{z})$ is the sum of squares $S^{(l)}$, considered as a function of λ and calculated from the standardized variable $z^{(\lambda)}$.

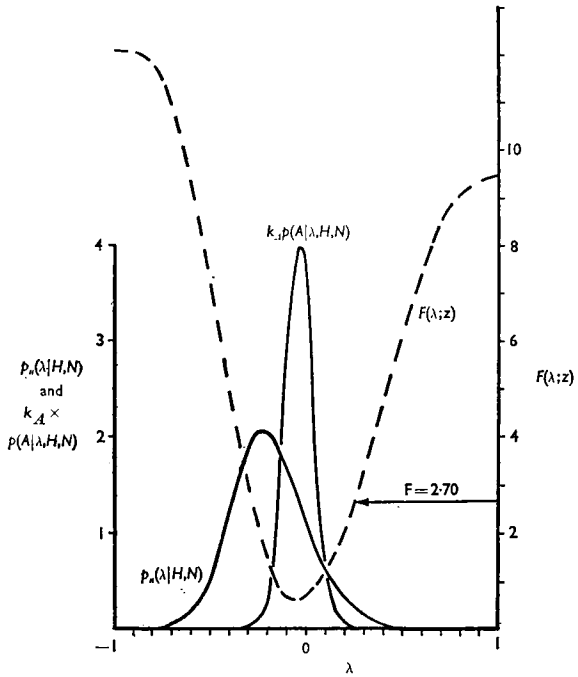

FIG. 5. Textile data. ——— Components of posterior distribution. – – – – – Variance ratio, $F(\lambda; \mathbf{z})$. Arrow gives 5 per cent significance level.

We now consider the constraint $H, \sigma_1^2 = \ldots = \sigma_k^2$, i.e. look at the possibility that a transformation exists simultaneously achieving normality and constant variance. Then if $S_\nu = \Sigma S^{(l)}$ is the pooled sum of squares within groups

$$L_{\max}(\lambda | H, N) = -\tfrac{1}{2} n \log\{S_\nu(\lambda; \mathbf{z})/n\}. \tag{46}$$

Therefore

$$L_{\max}(\lambda | H, N) = L_{\max}(\lambda | N) + \log\left[\frac{\Pi\{S^{(l)}(\lambda; \mathbf{z})/n_l\}^{\frac{1}{2}n_l}}{\{S_\nu(\lambda; \mathbf{z})/n\}^{\frac{1}{2}n}}\right]$$

$$= L_{\max}(\lambda | N) + \log L_1(\lambda; \mathbf{z}), \tag{47}$$

say. Here the second factor is the log of the Neyman–Pearson L_1 criterion for testing the hypothesis $\sigma_1^2 = \ldots = \sigma_k^2$.

In the corresponding Bayesian analysis, (29) gives

$$p(\lambda | H, N) = p(\lambda | N) \times k_H\, p(\sigma_1^2 = \ldots = \sigma_k^2 | \lambda, N), \tag{48}$$

where

$$k_H^{-1} = E_{\lambda|N}\{p(\sigma_1^2 = \ldots = \sigma_k^2 | \lambda, N)\}.$$

For the general model in which $\sigma_1^2, \ldots, \sigma_k^2$ may be different, the prior distribution is

$$p_0(\lambda)(\Pi d\theta_l)(\Pi d \log \sigma_l) J^{-\nu/n}$$

and

$$p(\lambda|N) = p_0(\lambda)\, c \Pi\{S^{(l)}(\lambda; \mathbf{z})\}^{-\frac{1}{2}\nu_l}, \tag{49}$$

with

$$c^{-1} = \int p_0(\lambda)\, \pi\{S^{(l)}(\lambda; \mathbf{z})\}^{-\frac{1}{2}\nu_l} d\lambda.$$

For the restricted model in which the variances are all equal to σ^2, the appropriate prior distribution is

$$p_0(\lambda)(\Pi d\theta_l)(d\log\sigma) J^{-\nu/n}$$

and

$$p(\lambda|H, N) = \{p_0(\lambda)\, c_\nu(\lambda; z)\}^{-\frac{1}{2}\nu}. \tag{50}$$

Hence, on dividing (50) by (49), we have that the second factor in (48) is

$$\frac{c_\nu}{c} \frac{\Pi\{S^{(l)}(\lambda; \mathbf{z})\}^{-\frac{1}{2}\nu_l}}{\{S_\nu(\lambda; \mathbf{z})\}^{-\frac{1}{2}\nu}} = \frac{c_\nu}{c} \frac{\Pi \nu_l^{\frac{1}{2}\nu_l}}{\nu^{\frac{1}{2}\nu}} e^{-\frac{1}{2}M(\lambda; \mathbf{z})}, \tag{51}$$

where (Bartlett, 1937)

$$M(\lambda; \mathbf{z}) = \nu \log\left\{\frac{S_\nu(\lambda; \mathbf{z})}{\nu}\right\} - \Sigma \nu_l \log\left\{\frac{S^{(l)}(\lambda; \mathbf{z})}{\nu_l}\right\}$$

is the modification of the L_1 statistic for testing homogeneity of variance, replacing sample sizes by degrees of freedom.

From our general argument, (51) must be proportional to $p(\sigma_1^2 = \ldots = \sigma_k^2 | \lambda, N)$. This can be verified directly by finding the joint posterior distribution of $\sigma_1^2, \ldots, \sigma_k^2$, transforming to new variables $\sigma_1^2, \sigma_2^2/\sigma_1^2, \ldots, \sigma_k^2/\sigma_1^2$, integrating out σ_1^2, and then taking unit values of the remaining arguments.

5.5. *Application to Biological Example*

In the biological example, we can now factorize the overall criterion into three parts. These correspond to the possibilities that in addition to normality within each group, we may be able to get constant variance and that it may be unnecessary to include interaction terms in the model, i.e. that additivity is achievable.

In terms of maximized likelihoods,

$$L_{\max}(\lambda|A, H, N) = L_{\max}(\lambda|N) + \log L_1(\lambda; \mathbf{z}) - \tfrac{1}{2} n \log\left\{1 + \frac{\nu_2}{\nu_r} F(\lambda; \mathbf{z})\right\}, \tag{52}$$

where $L_1(\lambda; \mathbf{z})$ is the criterion for testing constancy of variance given normality and $F(\lambda; \mathbf{z})$ is the criterion for absence of interaction given normality and constancy of variance.

The corresponding Bayesian analysis is

$$p(\lambda|A, H, N) = p(\lambda|N) \times k_H\, p(\sigma_1^2 = \ldots = \sigma_k^2 | \lambda, N) \times k_A\, p(\boldsymbol{\theta}_2 = 0 | \lambda, N, H). \tag{53}$$

The results are set out in Table 8 and in Figs. 6–8. The graphs of $L_{\max}(\lambda|N)$ and $p_u(\lambda|N)$ in Fig. 6 show that the information about λ coming from within group normality is very slight, values of λ as far apart as -1 and 2 being acceptable on this

basis. The requirement of constant variance, however, has a major effect on the choice of λ; further, some information is contributed by the requirement of additivity.

TABLE 8

Biological data. Calculations for analysis of the transformation

λ	$L_{\max}(\lambda \mid A, H, N)$	$L_{\max}(\lambda \mid H, N)$	$L_{\max}(\lambda \mid N)$	$M(\lambda; \mathbf{z})$	$F(\lambda; \mathbf{z})$
4·0			125·33		1·17
3·0			128·50		1·48
2·0	62·97	69·36	130·78	92·13	1·83
1·0	91·72	98·24	131·93	50·54	1·88
0·5	103·83	109·55	132·15	33·90	1·62
0·0	113·51	117·96	131·95	20·99	1·22
−0·2	116·44	120·37	131·79	17·13	1·07
−0·4	118·58	122·13	131·59	14·19	0·95
−0·6	119·82	123·21	131·35	12·21	0·90
−0·8	120·07	123·60	131·04	11·16	0·94
−1·0	119·29	123·30	130·69	11·09	1·09
−1·2	117·52	122·35	130·29	11·91	1·33
−1·4	114·86	120·76	129·85	13·64	1·67
−1·6	111·43	118·55	129·37	16·23	2·08
−2·0	102·74	112·50	128·27	23·66	3·01
−2·5	89·91	102·46	126·68	36·33	4·12
−3·0	75·69	90·10	124·84	52·11	4·93

λ	$p_u(\lambda \mid A, H, N)$	$p_u(\lambda \mid H, N)$	$p_u(\lambda \mid N)$	$k_H\, p(H \mid \lambda, N)$	$k_A\, p(A \mid \lambda, H, N)$
1·0			0·335		
0·5		0·006	0·398		0·03
0·0	0·006	0·021	0·342	0·06	0·28
−0·1	0·023	0·055	0·324	0·17	0·39
−0·2	0·076	0·127	0·304	0·42	0·60
−0·3	0·257	0·261	0·283	0·92	0·98
−0·4	0·492	0·471	0·261	1·80	1·04
−0·5	0·942	0·754	0·240	3·14	1·25
−0·6	1·462	1·059	0·218	4·85	1·38
−0·7	1·823	1·320	0·196	6·73	1·38
−0·8	1·823	1·430	0·173	8·27	1·27
−0·9	1·419	1·360	0·153	8·88	1·04
−1·0	0·923	1·136	0·134	8·47	0·81
−1·1	0·468	0·850	0·116	7·33	0·55
−1·2	0·194	0·558	0·099	5·64	0·35
−1·3	0·067	0·329	0·083	3·96	0·20
−1·4	0·019	0·170	0·069	2·46	0·11
−1·5	0·005	0·078	0·058	1·34	0·06
−1·6	0·001	0·032	0·050	0·64	0·03
−1·7		0·009			

From Fig. 7, which shows the detailed separation of the maximum-likelihood and Bayesian components, any transformation in the region y^{-1} to $y^{-\frac{1}{2}}$ gives a compatible compromise.

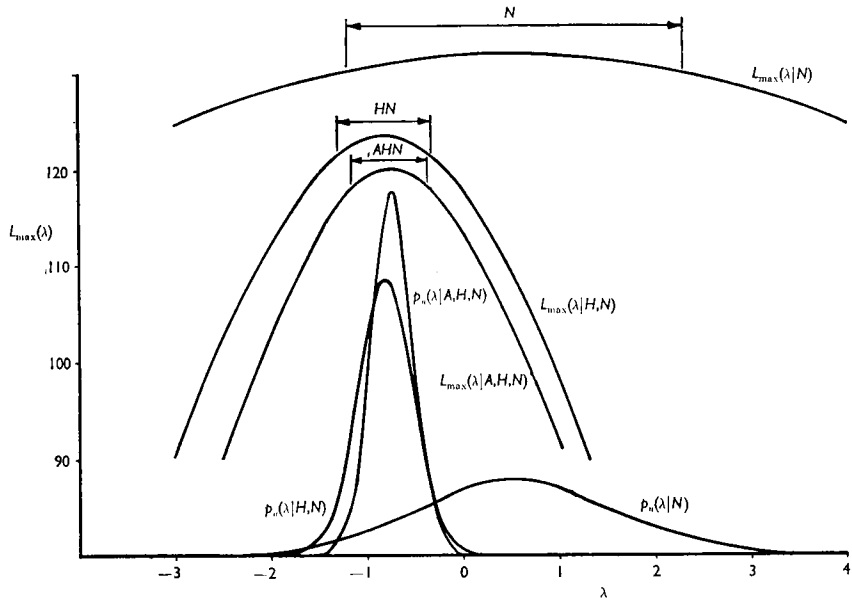

FIG. 6. Biological data. Functions $L_{\max}(\lambda)$ and $p_u(\lambda)$ under different models. *A*: additivity. *H*: homogeneity of variance. *N*: normality. Arrows *N*, *HN*, *AHN* show approximate 95 per cent confidence intervals for λ.

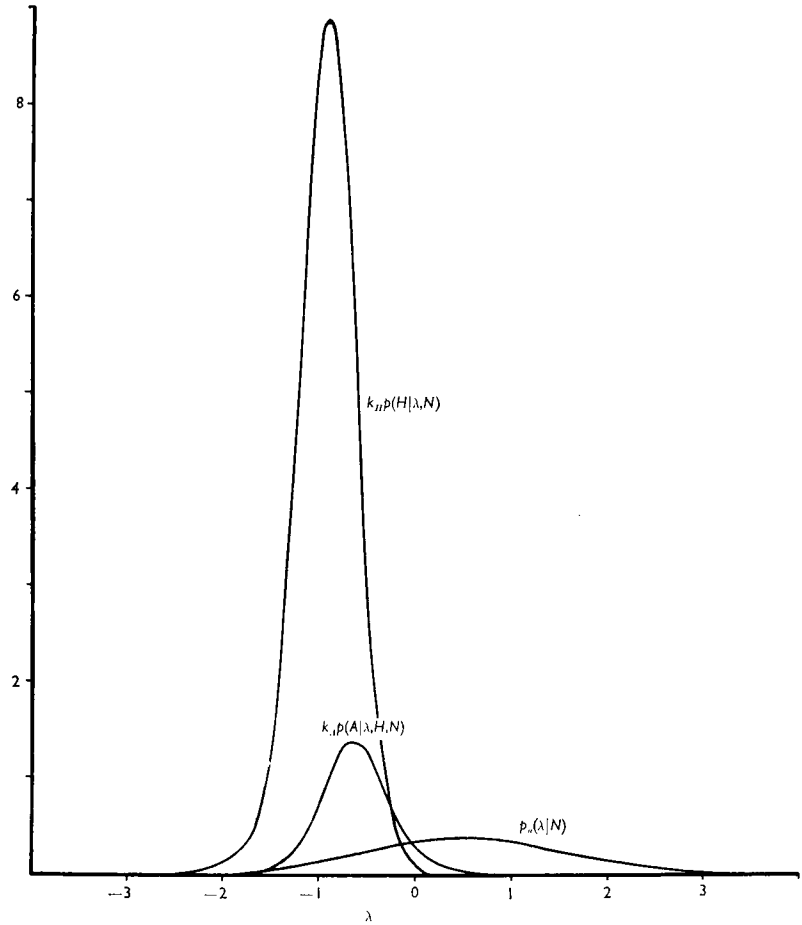

FIG. 7. Biological data. Components of posterior distribution.

Since the groups all contain four observations

$$-2\log L_1(\lambda; \mathbf{z}) = \tfrac{4}{3} M(\lambda; \mathbf{z})$$

and the graph of $M(\lambda; \mathbf{z})$ in Fig. 8 is equivalent to one of $L_1(\lambda; \mathbf{z})$. Since on the null hypothesis the distribution of $M(\lambda; \mathbf{z})$ is approximately χ^2_{11}, we can use Fig. 8 to

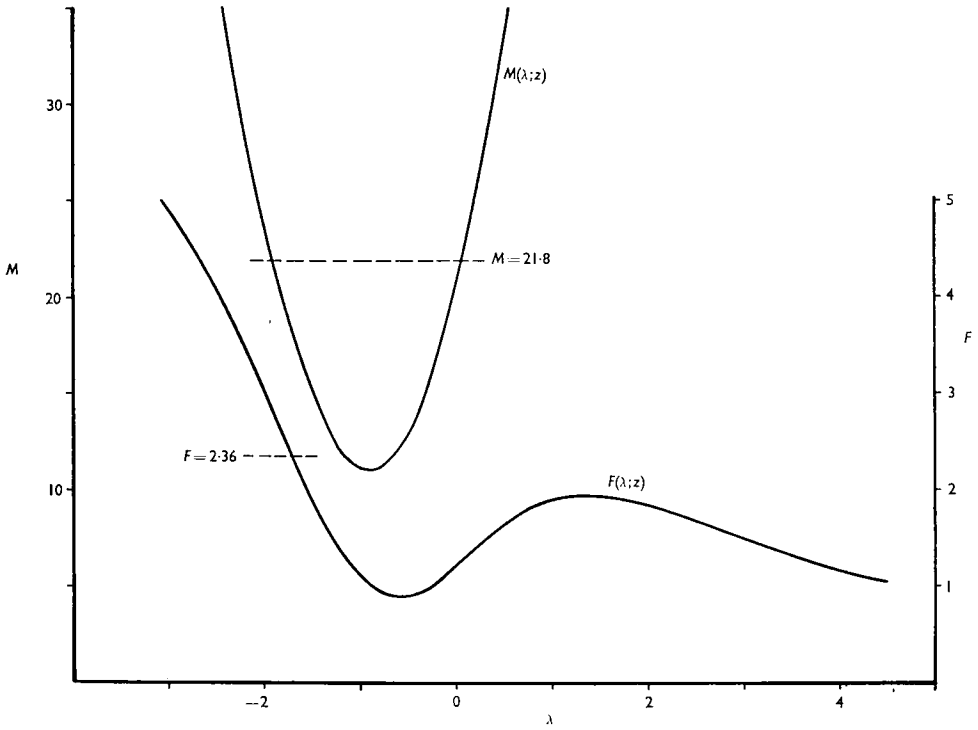

FIG. 8. Biological data. Variance ratio, $F(\lambda; \mathbf{z})$, for interaction against error as a function of λ. Bartlett's criterion, $M(\lambda; \mathbf{z})$, for equality of cell variances as a function of λ. Dotted lines give 5 per cent significance limits.

find the range in which the data are consistent with homoscedasticity. Similarly the graph of $F(\lambda; \mathbf{z})$ indicates the range within which the data are consistent with additivity. The dotted lines indicate the 5 per cent significance levels of M and of F.

The minimum of $M(\lambda; \mathbf{z})$ is very near $\lambda = -1$. It is of interest that the regression coefficient of log(sample variance) on log(sample mean) is nearly 4, so that the reciprocal transformation is suggested also by the usual approximate argument for stabilizing variance.

6. ANALYSIS OF RESIDUALS†

We now examine briefly a connection between the methods of the present paper and those based on the analysis of residuals. The analysis of residuals is intended

† We are greatly indebted to Professor F. J. Anscombe for pointing out an error in the approximation for α as we originally gave it. In the present modified version terms originally neglected in this Section have been included to correct the discrepancy.

primarily to examine what happens on one particular scale, although its use to indicate a transformation has been suggested (Anscombe and Tukey, 1963). Corresponding to an observation y, let Y be the deviation $\hat{y}-\bar{y}$ of the fitted value $\hat{y}$ from the sample mean and let $r = y-\hat{y}$ be the residual. If the ideal assumptions are satisfied r and Y will be distributed independently. Different sorts of departures from ideal assumptions can be measured, therefore, by studying the deviations of the statistics $T_{ij} = \Sigma r^i Y^j$ from $nE(r^i)E(Y^j)$. In addition to graphical analysis, a number of such functions have indeed been proposed for particular study (Anscombe, 1961; Anscombe and Tukey, 1963).

Specifically, the statistics

$$T_{30} = \Sigma r^3, \quad T_{40} = \Sigma r^4, \quad T_{21} = \Sigma r^2 Y, \quad T_{12} = \Sigma r Y^2 \tag{54}$$

were put forward as measures respectively of skewness, kurtosis, heterogeneity of variance and non-additivity. Tukey's degree of freedom for non-additivity (Tukey, 1949) involves the sum of squares corresponding to T_{12} considered as a contrast of residuals with "fixed" coefficients Y^2.

Suppose now that we consider the family of power transformations and, writing $z = y/\dot{y}$, and $w = z-1$, make the expansion

$$\begin{aligned} z^{(\lambda)} = \frac{z^\lambda - 1}{\lambda} &= w + \tfrac{1}{2}(\lambda-1)w^2 + \tfrac{1}{6}(\lambda-1)(\lambda-2)w^3 + O(w^4) \\ &= w - \alpha w_2 + \tfrac{2}{3}\alpha(\alpha+\tfrac{1}{2})w_3 + O(w^4), \end{aligned} \tag{55}$$

where $w_2 = w^2$, $w_3 = w^3$ and $\alpha = 1-\lambda$.

Now, $L_{\max}(\lambda)$ and $L_b(\lambda)$ are determined by the residual sum of squares of $z^{(\lambda)}$, which is approximately

$$\{\mathbf{w} - \alpha\mathbf{w}_2 + \tfrac{2}{3}\alpha(\alpha+\tfrac{1}{2})\mathbf{w}_3\}'\mathbf{a}_r\{\mathbf{w} - \alpha\mathbf{w}_2 + \tfrac{2}{3}\alpha(\alpha+\tfrac{1}{2})\mathbf{w}_3\}. \tag{56}$$

If we take terms up to the fourth degree in w and then differentiate with respect to α, we have that the maximum-likelihood estimate of α is approximately

$$\hat{\alpha} = \frac{3\mathbf{w}'\mathbf{a}_r\mathbf{w}_2 - \mathbf{w}'\mathbf{a}_r\mathbf{w}_3}{3\mathbf{w}_2'\mathbf{a}_r\mathbf{w}_2 + 4\mathbf{w}'\mathbf{a}_r\mathbf{w}_3}. \tag{57}$$

If we write $y_1 = y-\dot{y}$, $y_2 = (y-\dot{y})^2$, $y_3 = (y-\dot{y})^3$ and denote by $\hat{y}_1, \hat{y}_2, \hat{y}_3$ the values obtained by fitting y_1, y_2 and y_3 to the model, the above approximation may be expressed in terms of the original observations as

$$\hat{\alpha} = \frac{3\dot{y}(\mathbf{y}_1'\mathbf{a}_r\mathbf{y}_2) - \mathbf{y}_1'\mathbf{a}_r\mathbf{y}_3}{3\mathbf{y}_2'\mathbf{a}_r\mathbf{y}_2 + 4\mathbf{y}_1'\mathbf{a}_r\mathbf{y}_3} = \frac{3\dot{y}\Sigma(y_1-\hat{y}_1)(y_2-\hat{y}_2) - \Sigma(y_1-\hat{y}_1)(y_3-\hat{y}_3)}{3\Sigma(y_2-\hat{y}_2)^2 + 4\Sigma(y_1-\hat{y}_1)(y_3-\hat{y}_3)}. \tag{58}$$

To see the relation between this expression and the T statistics, write $d = \bar{y}-\dot{y}$. Then $y_1 = y-\dot{y} = r+Y+d$. Bearing in mind that $\mathbf{a}_r\mathbf{Y} = 0$, $\mathbf{a}_r\mathbf{r} = \mathbf{r}$, $\mathbf{Y}'\mathbf{r} = 0$, $\mathbf{a}_r\mathbf{1} = 0$, $\mathbf{1}'\mathbf{r} = 0$, where $\mathbf{1}$ denotes a vector of ones, terms such as $\mathbf{y}_1'\mathbf{a}_r\mathbf{y}_2$ can easily be expressed in terms of sums of powers and products of r, Y and d. In particular, on writing S for Σr^2, we find the numerator of (58) to be

$$3(\bar{y}-3d)(T_{30}+2T_{21}+T_{12}) - (T_{40}+3T_{31}+3T_{22}+T_{13}) + 3d(2\bar{y}-3d)S. \tag{59}$$

To this order of approximation the maximum-likelihood estimate of α thus involves all the T statistics of orders 3 and 4.

As a very special case, for data assumed to form a single random sample

$$\hat{\alpha} = \frac{3\bar{y}\Sigma(y-\bar{y})(y_2-\bar{y}_2)+\Sigma(y-\bar{y})(y_3-\bar{y}_3)}{3\Sigma(y_2-\bar{y}_2)^2+4\Sigma(y-\bar{y})(y_3-\bar{y}_3)}.$$

Here questions such as non-additivity and non-constancy of variance do not arise and the transformation is attempting only to produce normality. Correspondingly in (59), $T_{21} = T_{12} = T_{31} = T_{22} = T_{13} = 0$, since $Y = \hat{y} - \bar{y} = 0$. In fact if we write $m_1 = \bar{y}$, $m_p = n^{-1}\Sigma(y-\bar{y})^p$ $(p = 2, 3, \ldots)$ and make the approximation $d = \frac{1}{2}m_2/m_1$, we have that

$$\hat{\alpha} = \frac{m_1 m_3 - \frac{1}{3}\left\{(m_4 - 3m_2^2) + \dfrac{3m_2 m_3}{m_1} + \dfrac{9}{4}\dfrac{m_2^3}{m_1^2}\right\}}{6m_2^2 + \frac{1}{3}\left\{7(m_4 - 3m_2^2) + 12\dfrac{m_2 m_3}{m_1} + 6\dfrac{m_2^3}{m_1^2}\right\}}. \tag{60}$$

For distributions in which m_1, m_2, m_3 and $m_4 - 3m_2^2$ are of the same order of magnitude, the terms in curly brackets are of one order higher in $1/m_1$ than are the other terms of the numerator and denominator. If we ignore the higher-order terms, we have

$$\hat{\alpha} \simeq \frac{m_1 m_3}{6m_2^2}.$$

A useful check suggested by Anscombe is to consider the χ^2 distribution for moderate degrees of freedom and the Poisson distribution for not too small a mean. For χ^2 we find $\alpha \simeq \frac{1}{6}$, whence $\lambda \simeq \frac{1}{3}$, corresponding to the well-known Wilson–Hilferty transformation. For the Poisson distribution, $\alpha \simeq \frac{1}{3}$, whence $\lambda \simeq \frac{2}{3}$.

7. Analysis of Effects after Transformation

In Section 2 we suggested that, having chosen a suitable λ, we should make the usual detailed estimation and interpretation of effects on this transformed scale. Thus in our two examples we recommended that the detailed interpretation should be in terms of a standard analysis of respectively $1/y$ and $\log y$. Since the value of λ used is selected at least partly in the light of the data, the question arises of a possible need to allow for this selection when interpreting the factor effects.

To investigate an appropriate allowance, we regard λ as an unknown parameter with "true" value λ_0, say, and suppose the true factor effects to be measured in terms of the scale λ_0. If we were, for instance, to analyse the factor effects on the scale corresponding to the maximum-likelihood estimate $\hat{\lambda}$, we might expect some additional error arising from the difference between $\hat{\lambda}$ and λ_0. We now investigate this matter, although the present formulation of the problem is not always completely realistic. For example, in our biological example, having decided to work with $1/y$, we shall probably be interested in factor effects measured on this scale and not those measured in some unknown scale corresponding to an unknown "true" λ_0. On the other hand, if we are interested in whether there is interaction between two factors, it is possibly dangerous to answer this by testing for interaction on the scale $\hat{\lambda}$, since $\hat{\lambda}$ may be selected at least in part to minimize the sample interaction. A more reasonable formulation here may often be: on some unknown "true" scale λ_0, are interaction terms necessary in the model?

From the maximum-likelihood approach, the most useful result is that significance tests for null hypotheses, such as that just mentioned about the absence of interaction, can be obtained in a straightforward way in terms of the usual large-sample chi-squared test. Thus, in the textile example, we could test the null hypothesis that second-degree terms are absent for some unknown "true" λ_0, by testing twice the difference of the maxima of the two curves of $L_{\max}(\lambda)$ in Fig. 4 as χ^2_6. Note that the maxima occur at different values of λ. In this particular example, such a test is hardly necessary.

It would be possible to obtain more detailed results by evaluating the usual large-sample information matrix for the joint estimation of λ, σ^2 and $\boldsymbol{\theta}$. Since, however, more specific results can be obtained from the Bayesian analysis, we shall present only those. The general conclusion will be that to allow for the effect of analysing in terms of $\hat{\lambda}$ rather than λ_0, the residual degrees of freedom need only be reduced by ν_λ, the number of component parameters in λ. This result applies provided that the population and sample effects are measured in terms of the normalized variables $\mathbf{z}^{(\lambda)}$.

Consider locally uniform prior densities for $\boldsymbol{\theta}$, $\log\sigma$ and λ. Then the posterior density for $\boldsymbol{\theta}$ is

$$\frac{\int\{(\mathbf{z}^{(\lambda)}-\mathbf{a}\boldsymbol{\theta})'(\mathbf{z}^{(\lambda)}-\mathbf{a}\boldsymbol{\theta})\}^{-\frac{1}{2}n}d\lambda}{\int\{\nu_r s^2(\lambda;\mathbf{z})\}^{-\frac{1}{2}\nu_r}d\lambda}. \tag{61}$$

Approximate evaluation of the integral in (61) is done by expansion around the maxima of the integrands. The maximum of the integrand in the denominator is at the maximum-likelihood estimate $\hat{\lambda}$, and that of the numerator is near $\hat{\lambda}$, so long as $\boldsymbol{\theta}$ is near its maximum-likelihood value. The answer is that (61) is approximately

$$\frac{\{(\mathbf{z}^{(\hat{\lambda})}-\mathbf{a}\boldsymbol{\theta})'(\mathbf{z}^{(\hat{\lambda})}-\mathbf{a}\boldsymbol{\theta})\}^{-\frac{1}{2}(n-\nu_\lambda)}}{\{\nu_r s^2(\hat{\lambda};\mathbf{z})\}^{-\frac{1}{2}(\nu_r-\nu_\lambda)}}. \tag{62}$$

This is exactly the posterior density of $\boldsymbol{\theta}$ for some known fixed λ with the degrees of freedom reduced by ν_λ.

To derive (62) from (61), we need to evaluate integrals of the form

$$I=\int\{q(\lambda)\}^{-\frac{1}{2}\nu}d\lambda, \tag{63}$$

where ν is large, and $q(\lambda)$ is assumed positive and to have a unique minimum at $\lambda=\hat{\lambda}$, with a finite Hessian determinant Δ_q at the minimum. We can then make a Laplace expansion, writing

$$\begin{aligned} I&=\int\exp\left[-\frac{\nu}{2}\log q(\hat{\lambda})-\frac{\nu}{2}\log\left\{1+\frac{q(\lambda)-q(\hat{\lambda})}{q(\hat{\lambda})}\right\}\right]d\lambda \\ &\simeq\frac{\{q(\hat{\lambda})\}^{-\frac{1}{2}\nu-\frac{1}{2}\nu_\lambda}}{\Delta_q^{\frac{1}{2}}}\times\text{const}; \end{aligned} \tag{64}$$

for this we expand the second logarithmic term as far as the quadratic terms and then integrate over the whole ν_λ-dimensional space of λ. In our application the terms $\Delta_q^{\frac{1}{2}}$ in numerator and denominator are equal to the first order.

Finally, we can obtain an approximation to the posterior distribution $p_u(\boldsymbol{\lambda})$ of $\boldsymbol{\lambda}$ that is better than the usual type of asymptotic normal approximation. For an expansion about $\boldsymbol{\lambda}$ gives that

$$p_u(\boldsymbol{\lambda}) = \frac{\{s^2(\boldsymbol{\lambda}; \mathbf{z})\}^{-\frac{1}{2}\nu_r}}{\int\{s^2(\boldsymbol{\lambda}; \mathbf{z})\}^{-\frac{1}{2}\nu_r} d\boldsymbol{\lambda}}$$

$$\simeq \frac{\text{const}}{\left\{1+\dfrac{(\boldsymbol{\lambda}-\hat{\boldsymbol{\lambda}})'\mathbf{b}(\boldsymbol{\lambda}-\hat{\boldsymbol{\lambda}})}{\nu_r s^2(\hat{\boldsymbol{\lambda}}; \mathbf{z})}\right\}^{\frac{1}{2}\nu_\lambda}}. \tag{65}$$

Here

$$\mathbf{b} = \mathbf{d}'(\hat{\boldsymbol{\lambda}})\,\mathbf{a}_r\,d(\hat{\boldsymbol{\lambda}}), \tag{66}$$

with $\mathbf{d}(\boldsymbol{\lambda})$ being the $n \times \nu_\lambda$ matrix with elements

$$\frac{\partial z_i^{(\lambda)}}{\partial \lambda_j} \quad (i = 1, \ldots, n;\ j = 1, \ldots, \nu_\lambda).$$

The matrix $\mathbf{b}$ determines the quadratic terms in the expansion of $s^2(\boldsymbol{\lambda}; \mathbf{z})$ around $\hat{\boldsymbol{\lambda}}$.

Thus the quantities $(\lambda_j - \hat{\lambda}_j)/\{s(\hat{\boldsymbol{\lambda}}; \mathbf{z})\sqrt{b^{ii}}\}$ have approximately a posterior multivariate t distribution and

$$\frac{(\boldsymbol{\lambda}-\hat{\boldsymbol{\lambda}})'\mathbf{b}(\boldsymbol{\lambda}-\hat{\boldsymbol{\lambda}})}{\nu_r s^2(\hat{\boldsymbol{\lambda}}; \mathbf{z})}$$

a posterior F distribution. In fact, however, it will usually be better to examine the posterior distribution of $\boldsymbol{\lambda}$ directly, as we have done in the numerical examples.

8. FURTHER DEVELOPMENTS

We now consider in much less detail a number of possible developments of the methods proposed in this paper. Of these, the most important is probably the simultaneous transformation of independent and dependent variables in a regression problem. Some general remarks on this have been made in Section 1.

Denote the dependent variable by y and the independent variables by $x_1, \ldots, x_l$. Consider a family of transformations from y into $y^{(\lambda)}$ and $x_1, \ldots, x_l$ into $x_1^{(\kappa_1)}, \ldots, x_l^{(\kappa_l)}$, the whole transformation being thus indexed by the parameters $(\lambda; \kappa_1, \ldots, \kappa_l)$. It is not necessary that the family of transformations of say x_1 into $x_1^{(\kappa_1)}$ and x_2 into $x_2^{(\kappa_2)}$ should be the same, although this would often be the case.

We now assume that for some unknown $(\lambda; \kappa_1, \ldots, \kappa_l)$ the usual normal theory assumptions of linear regression theory hold. We can then compute say the maximized log likelihood for given $(\lambda; \kappa_1, \ldots, \kappa_l)$, obtaining exactly as in (8)

$$L_{\max}(\lambda; \kappa_1, \ldots, \kappa_l) = -\tfrac{1}{2}\log\hat{\sigma}^2(\lambda; \kappa_1, \ldots, \kappa_l) + \log J(\lambda; \mathbf{y}), \tag{67}$$

where $\hat{\sigma}^2(\lambda; \kappa_1, \ldots, \kappa_l)$ is the maximum-likelihood estimate of residual variance in the standard multiple regression analysis of the transformed variable. The corresponding expression from the Bayesian approach is

$$L_b(\lambda; \kappa_1, \ldots, \kappa_l) = -\tfrac{1}{2}\nu_r \log s^2(\lambda; \kappa_1, \ldots, \kappa_l) + \frac{\nu_r}{n}\log J(\lambda; \mathbf{y}). \tag{68}$$

The straightforward extension of the procedure of Section 3 is to compute (67) or (68) for a suitable set of $(\lambda; \kappa_1, ..., \kappa_l)$ and to examine the resulting surface especially near its maximum. This is, however, a tedious procedure, except perhaps for $l = 1$. Further, graphical presentation of the conclusions will not be easy if $l > 1$; for $l = 1$ we can plot contours of the functions (67) and (68).

When λ is fixed, i.e. transformations of the independent variables only are involved, Box and Tidwell (1962) developed an iterative procedure for the corresponding non-linear least-squares problem. In this the independent variables are, if necessary, first transformed to near the optimum form. Then two terms of the Taylor expansion of $x_1^{(\kappa_1)}, ..., x_l^{(\kappa_l)}$ are taken. For example if $x_1^{(\kappa_1)} = x^{\kappa_1}$ and the best value for κ_1 is thought to be near 1, we write

$$x_1^{\kappa_1} = x_1 + (\kappa_1 - 1) x_1 \log x_1. \tag{69}$$

A linear regression term $\beta_1 x_1^{\kappa_1}$ can then be written approximately

$$\beta_1 x_1 + \beta_1(\kappa_1 - 1) x_1 \log x_1 = \beta_1 x_1 + \gamma_1 x_1 \log x_1,$$

say. If the linear model involves linear regression on $x_1, ..., x_l$ and if all the transformations of the independent variable are to powers, we can therefore take the linear regression on $x_1, ..., x_l, x_1 \log x_1, ..., x_l \log x_l$ in order to estimate the β's and γ's and hence also the κ's. The procedure can then be iterated. Transformation of the dependent variable will usually be the more critical. Therefore, a reasonable practical procedure will often be to combine straightforward investigation of transformation of the dependent variable with Box and Tidwell's method applied to the independent variables.

It is possible also to consider simplifications of the procedure for determining a transformation of the dependent variable. The main labour in straightforward application of the method of Section 3 is in applying the transformation for various values of λ and then computing the standard analysis of variance for each set of transformed data. Such a sequence of similar calculations is straightforward on an electronic computer. It is perfectly practicable also for occasional desk calculation, although probably not for routine use. There are a number of possible simplifications based, for example, on expansions like (69) or even (55), but they have to be used very cautiously.

In the present paper we have concentrated largely on transformations for those standard "fixed-effects" analysis of variance situations where the response can be treated as a continuous variable. The same general approach could be adopted in dealing with "random-effects" models, and with various problems in multivariate analysis and in the analysis of time series. We shall not go into these applications here.

An important omission from our discussion concerns transformations specifically for data suspected of following the Poisson or binomial distributions. There are two difficulties here. One is purely computational. Suppose we assume that our observations, y, follow, for example, Poisson distributions with means that obey an additive law on an unknown transformed scale. Thus, in a row–column arrangement, it might be assumed that the Poisson mean in row i and column j has the form

$$(\mu + \alpha_i + \beta_j)^{1/\lambda} \quad (\lambda \neq 0),$$

$$\mu \alpha_i \beta_j \quad (\lambda = 0),$$

where λ is unknown. Then λ and the other parameters of the model can be estimated by maximum likelihood (Cochran, 1940). It would probably be possible to develop reasonable approximations to this procedure although we have not investigated this matter.

An essential distinction between this situation and the one considered in Section 3 is that here the untransformed observations y have known distributional properties. The analogous normal theory situation would involve observations y normally distributed with constant variance on the untransformed scale, but for which the population means are additive on a transformed scale. The maximum-likelihood solution in this case would involve, at least in principle, a straightforward non-linear least-squares problem. However, this situation does not seem likely to arise often; certainly, it is inappropriate in our examples.

An important possible complication of the analysis of data connected with Poisson and binomial distributions has been particularly stressed by Bartlett (1947). This is the presence of an additional component of variance of unknown form on top of the Poisson or binomial variation. If inspection of the data shows that such additional variation is substantial, it may be adequate to apply the methods of Section 3. For integer data with range $(0, 1, \ldots)$ it will often be reasonable to consider power transformations. For data in the form of proportions of "successes" in which "successes" and "failures" are to be treated symmetrically, Professor J. W. Tukey has, in an unpublished paper, suggested the family of transformations from y to

$$y^{\lambda} - (1-y)^{\lambda}.$$

For suitable λ's this approximates closely to the standard transforms of proportions, the probit, logistic and angular transformations. The methods of the present paper could be applied with this family of transformations.

ACKNOWLEDGEMENT

We thank many friends for remarks leading to the writing of this paper.

REFERENCES

ANSCOMBE, F. J. (1961), "Examination of residuals", *Proc. Fourth Berkeley Symp. Math. Statist. and Prob.*, **1**, 1–36.

—— and TUKEY, J. W. (1963), "The examination and analysis of residuals", *Technometrics*, **5**, 141–160.

BARTLETT, M. S. (1937), "Properties of sufficiency and statistical tests", *Proc. Roy. Soc.* A, **160**, 268–282.

—— (1947), "The use of transformations", *Biometrics*, **3**, 39–52.

BOX, G. E. P. and TIDWELL, P. W. (1962), "Transformation of the independent variables", *Technometrics*, **4**, 531–550.

COCHRAN, W. G. (1940), "The analysis of variance when experimental errors follow the Poisson or binomial laws", *Ann. math. Statist.*, **11**, 335–347.

CORNISH, E. A. (1954), "The multivariate t distribution associated with a set of normal sample deviates", *Austral. J. Physics*, **7**, 531–542.

DUNNETT, C. W. and SOBEL, M. (1954), "A bivariate generalization of Student's t distribution", *Biometrika*, **41**, 153–169.

JEFFREYS, H. (1961), *Theory of Probability*, 3rd ed. Oxford University Press.

KLECZKOWSKI, A. (1949), "The transformation of local lesion counts for statistical analysis", *Ann. appl. Biol.*, **36**, 139–152.

TUKEY, J. W. (1949), "One degree of freedom for non-additivity", *Biometrics*, **5**, 232–242.

—— (1950), "Dyadic anova, an analysis of variance for vectors", *Human Biology*, **21**, 65–110.

—— and MOORE, P. G. (1954), "Answer to query 112", *Biometrics*, **10**, 562–568.