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 $y_1, y_2, ..., y_n$ are independently normally distributed with constant variance and with expectations specified by a model linear in a set of parameters $\boldsymbol{\theta}$. 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 $x_1, x_2, ..., x_l$ 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

$$y \propto x_1^{\beta 1} \dots x_l^{\beta l}$$

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

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

and hence, to this order, the prior density of σ_λ^2 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

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

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

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

The integral (21) can be evaluated to give

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

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 $L_{\text{max}}(\lambda)$ 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 $L_{\text{max}}(\lambda)$ 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		

$L_{\max}(\lambda) = -24\log\hat{\sigma}^2(\lambda;\mathbf{z}) = \log\{S(\lambda;\mathbf{z})\}^{-24}+92{\cdot}91;\ p_u(\lambda) = k\,e^{L_b(\lambda)} = 0{\cdot}866\times10^{-10}\{S(\lambda;\mathbf{z})\}^{-21}.$

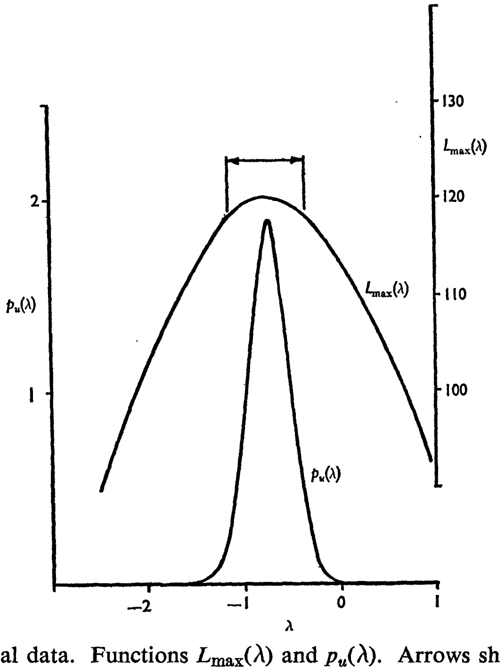

FIG. 1. Biological data. Functions $L_{\max}(\lambda)$ and $p_u(\lambda)$. Arrows show approximate 95 per cent. confidence interval for λ.

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 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.

TABLE 3

Analyses of variance of biological data

	Degrees of freedom	*Mean squares* × 1000 *Untransformed*	*Mean squares* × 1000 *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

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

$$y \propto x_1^{\beta_1} x_2^{\beta_2} x_3^{\beta_3}. \qquad (25)$$

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 example, the approximate 95 per cent confidence range extending only from $-0{\cdot}18$ to $+0{\cdot}06$. The posterior distribution $p_u(\lambda)$ has its mean at $-0{\cdot}06$. About 95 per cent of the distribution is included between $-0{\cdot}20$ and $+0{\cdot}08$. 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.

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

$L_{\max}(\lambda) = -13{\cdot}5 \log \hat{\sigma}^2(\lambda; \mathbf{z}) = \{S(\lambda; \mathbf{z})\}^{-13\cdot5} + 44{\cdot}49.$

$p_u(\lambda) = k\, e^{L_b(\lambda)} = 0{\cdot}540 \times 10^{-6} \{S(\lambda; \mathbf{z})\}^{-11\cdot5}.$

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

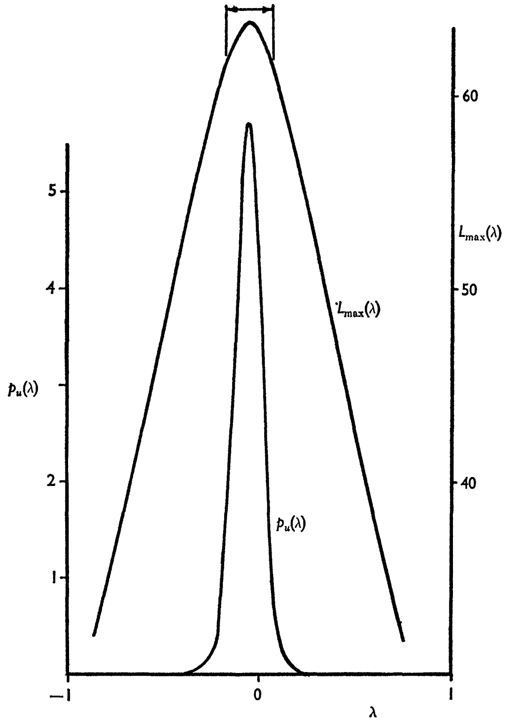

Fig. 2. Textile data. Functions $L_{max}(\lambda)$ and $p_u(\lambda)$. Arrows show approximate 95 per cent confidence interval for λ.

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

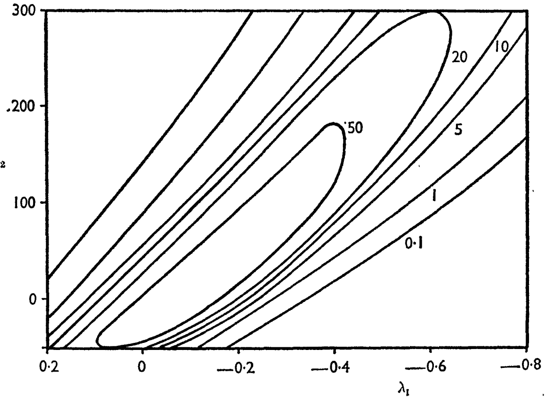

FIG. 3. Textile data. Transformation to $(y+\lambda_2)^{\lambda_1}$. Contours of $p_u(\lambda_1, \lambda_2)$ labelled with approximate percentage of posterior distribution excluded.

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

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

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

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

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

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

where $p(C) = E_\lambda\{p(C \mid \lambda)\}$ 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,

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

where $p(C_2 \mid C_1) = E_\lambda\{p(C_2 \mid \lambda, C_1)\}$ 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 $p(\lambda \mid C_1, C_2)$.

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,

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

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

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

whereas

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

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

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

where

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

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

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

where

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

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

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

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

where the normalizing constant is given by

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

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

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

with

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

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 $p(\boldsymbol{\theta}_2 = 0 \mid \lambda, H, N)$. 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 $\log \sigma$, 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

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

whence

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

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

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

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

thus gives the much wider approximate 95 per cent confidence interval (−0·48, 0·13) 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

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

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* $= -13{\cdot}5 \times$	
			$\log(1+\frac{6}{17}F(\lambda;\, z))$	$F(\lambda;\, z)$
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.

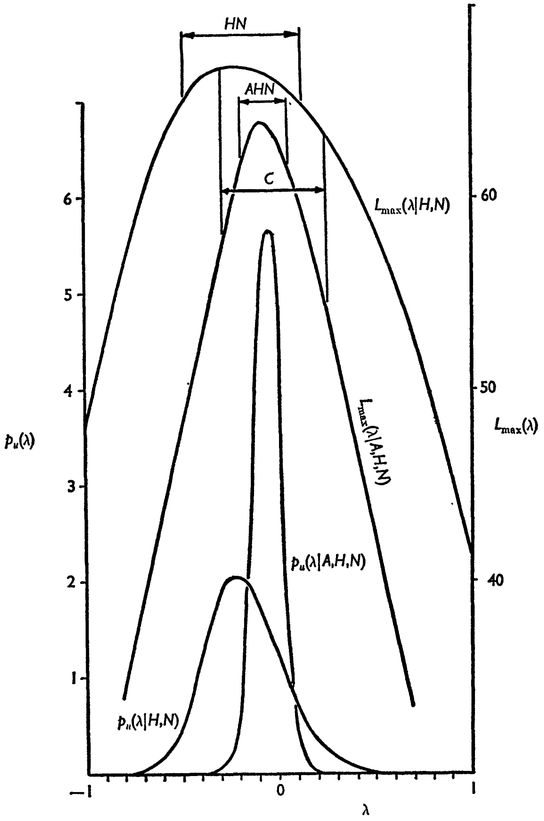

FIG. 4. Textile data. Functions $L_{\max}(\lambda)$ and $p_u(\lambda)$ under different models. A: additivity. H: homogeneity of variance. N: normality. Arrows *HN, AHN* show approximate 95 per cent confidence intervals for λ. Arrows C show range for which F for second-degree terms is not significant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

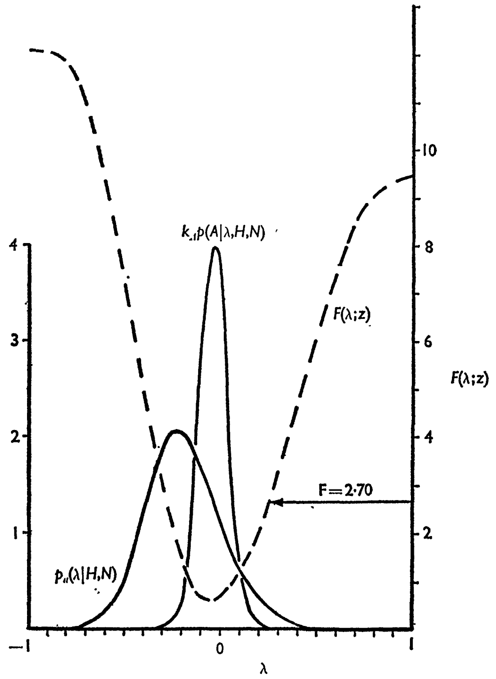

FIG. 5. Textile data. ——— Components of posterior distribution. – – – – – Variance ratio, $F(\lambda; \mathbf{z})$. Arrow gives 5 per cent significance level.

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

$$\begin{aligned} L_{\max}(\lambda | H, N) &= L_{\max}(\lambda | N) + \log\left[\frac{\Pi\{S^{(l)}(\lambda; \mathbf{z})/n_l\}^{\frac{1}{2}n_l}}{\{S_\nu(\lambda; \mathbf{z})/n\}^{\frac{1}{2}n}}\right] \\ &= L_{\max}(\lambda | N) + \log L_1(\lambda; \mathbf{z}), \end{aligned} \tag{47}$$

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,

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

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.

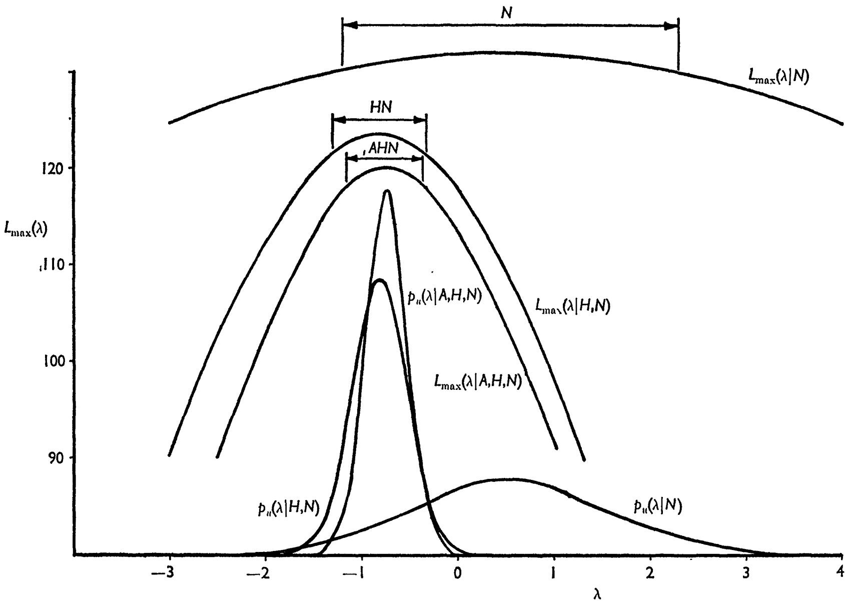

FIG. 6. Biological data. Functions $L_{\max}(\lambda)$ and $p_u(\lambda)$ under different models. A: additivity. H: homogeneity of variance. N: normality. Arrows N, HN, AHN show approximate 95 per cent confidence intervals for λ.

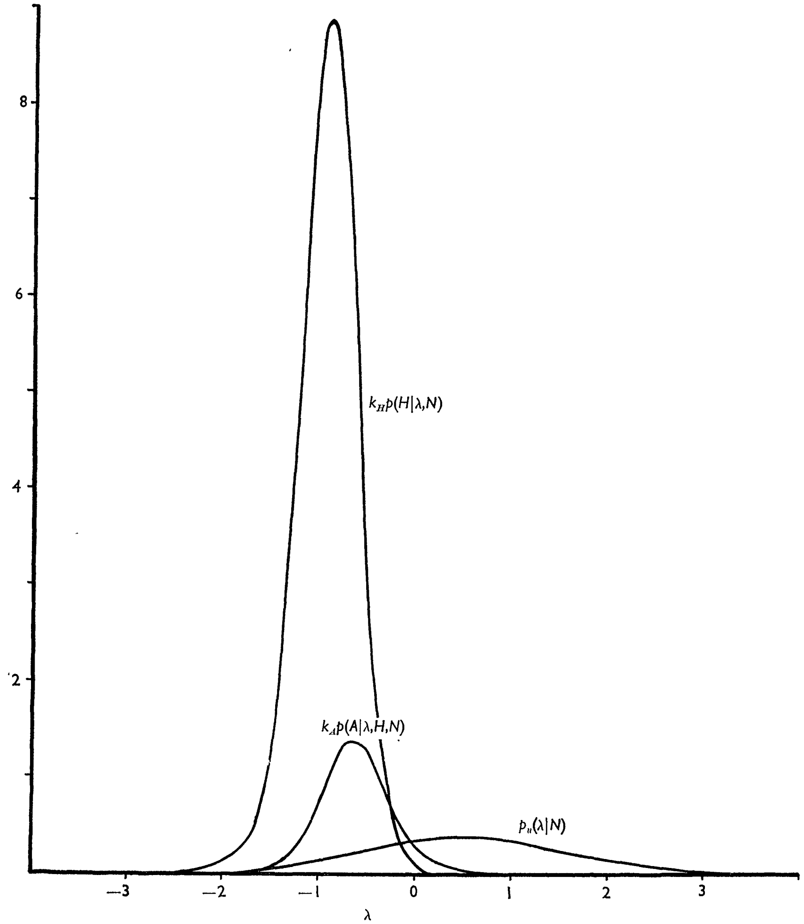

FIG. 7. Biological data. Components of posterior distribution.

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 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.

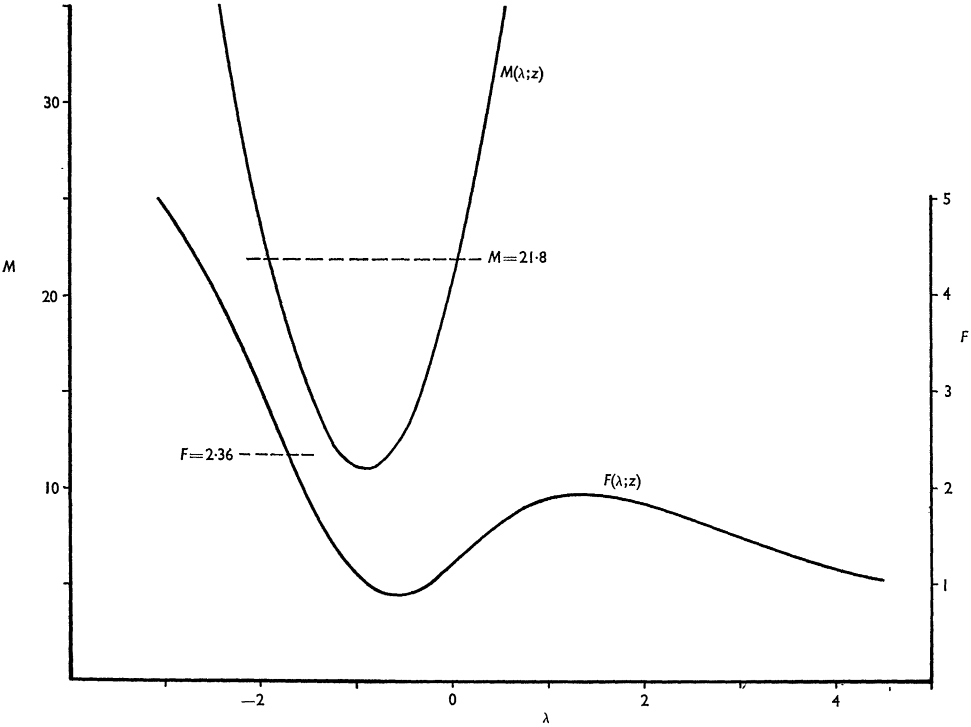

Fig. 8. Biological data. Variance ratio, $F(\lambda;\, z)$, for interaction against error as a function of λ. Bartlett's criterion, $M(\lambda;\, z)$, for equality of cell variances as a function of λ. Dotted lines give 5 per cent significance limits.

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

$$\hat{\alpha}=\frac{m_1m_3-\dfrac{1}{3}\left\{(m_4-3m_2^2)+\dfrac{3m_2m_3}{m_1}+\dfrac{9}{4}\dfrac{m_2^3}{m_1^2}\right\}}{6m_2^2+\dfrac{1}{3}\left\{7(m_4-3m_2^2)+12\dfrac{m_2m_3}{m_1}+6\dfrac{m_2^3}{m_1^2}\right\}}. \qquad (60)$$

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

$$\hat{\alpha}\simeq\frac{m_1m_3}{6m_2^2}.$$

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_6^2. 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

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

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

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

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

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

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

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

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

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

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

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 $(\lambda;\ \kappa_1, \ldots, \kappa_l)$ 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 $x_1^{(\kappa_1)}, \ldots, x_l^{(\kappa_l)}$ 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 $x_1, \ldots, x_l$ and if all the transformations of the independent variable are to powers, we can therefore take the linear regression on $x_1, \ldots, x_l$, $x_1 \log x_1, \ldots, x_l \log x_l$ 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

$$\begin{aligned} &(\mu + \alpha_i + \beta_j)^{1/\lambda} && (\lambda \neq 0), \\ &\mu \alpha_i \beta_j && (\lambda = 0), \end{aligned}$$

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.

Discussion on Paper by Professor Box and Professor Cox

Mr J. A. Nelder: May I begin with a definition (from the Concise Oxford Dictionary): "Box and Cox—two persons who take turns in sustaining a part." I must admit to having spent some time in trying to deduce which person was sustaining which part of this most interesting paper. I do not think the exercise was very successful, and this testifies to some sound collaboration on the part of the authors.

It seems to me that there are two basic problems besetting all conscientious data analysts (to borrow Professor Tukey's term). One is how to check that the data are not contaminated with rogue observations and what action to take if they are. The other is how to check that the model being used to analyse the data is substantially the right one. Looking through the corpus of statistical writings one must be struck, I think, by how relatively little effort has been devoted to these problems. The overwhelming preponderance of the literature consists of deductive exercises from *a priori* starting points. Now, of course, there must always be some assumptions made *a priori*; in data analysis the important thing is that they should not be much stronger than previous evidence justifies. The first of the two problems, that of gross errors or rogue observations, we are not directly concerned with now, but the question of scale for analysis, which is discussed here, is fundamental to the second. One sees not infrequently remarks to the effect that the design of an experiment determines the analysis. Life would be easier if this were true. To the information from the design we must add the analyst's prior judgements, preconceptions or prejudices (call them what you will) about questions of additivity, homoscedasticity and the like. Frequently these prior assumptions are unjustifiably strong, and amount to an assertion that the scale adopted will give the required additivity, etc. The great virtue of this paper lies in its showing us how to weaken these prior assumptions and allow the data to speak for themselves in these matters. The data analyst's two problems are closely intertwined, however; for if rogue observations are present their residuals tend to dominate the residual sum of squares, and may thus seriously affect the estimation of λ.

The two approaches, via likelihood and via Bayes theorem, run side by side, and give results which will often be very similar. I am not entirely happy about the derivation of equation (19) and wonder whether the appearance of the observations in the prior probability is not only "interesting", as the authors state, but also illegal. They remark (on p. 219) that, "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." I agree and, furthermore, I believe that a suitable modification of the likelihood approach may be found to produce just this result. The starting point is that fixed effects are unrealistic in a model. If we measure a treatment effect in an experiment, it is common experience that a further experiment will give us a further estimate of the effect which often differs from the original estimate by more than the internal standard errors of the experiments would lead us to expect. If we construct a model with this in mind, then for a single normal sample of n we might obtain

$$y_i = m + e_i$$

where $m = N(\mu, \sigma'^2)$ and $e_i = N(0, \sigma^2)$. If we now do an orthogonal transformation of the data $\mathbf{z} = \mathbf{Hy}$ where $\mathbf{H}$ is an orthogonal matrix of known coefficients having its first row with elements $n^{-\frac{1}{2}}$, then the log likelihood is given by

$$L = \text{const} - \tfrac{1}{2} \ln V - (z_1 - \mu\sqrt{n})^2/2V - \tfrac{1}{2}(n-1) \log \sigma^2 - \sum_2^n z_i^2/2\sigma^2,$$

where $V = \sigma^2 + n\sigma'^2$. Clearly we cannot estimate V unless μ is known, which in general it is not. However, for any fixed *but unknown* V, we have L maximized by taking

$$\hat{\mu} = \bar{y}, \quad \text{and} \quad \hat{\sigma}^2 = \Sigma(y - \bar{y})^2/(n-1).$$

Thus $L_{\max}(\lambda)$ following equation (24) is replaced (apart from an unknown constant) by $L_b(\lambda)$. By extensions of this argument we obtain Bartlett's criterion for testing the homogeneity of variances instead of the L_1 criterion, and the likelihood criterion for a restricted hypothesis on the means (equation (35)) becomes the same (apart from an unknown constant factor) as the Bayesian one. Thus some of the apparent differences between the two approaches may result from the restrictions implied by fixed effects in a model, these being equivalent to assertions of zero variance in repetitions of the experiment.

Taken with the work of Tukey, Daniel and others, on the detection of rogue observations, the results of this paper should lead before long to substantial improvements in computer programmes for the analysis of experiments. "First generation" programmes, which largely behave as though the design *did* wholly define the analysis, will be replaced by new second-generation programmes capable of checking the additional assumptions and taking appropriate action. It is hardly necessary to stress what an advance this would be.

I suppose that the converse of "two persons who take turns in sustaining a part" would be "one person who takes turns in sustaining two parts". Such a person is often the proposer of the vote of thanks, the parts being those of congratulator and critic; the latter has been known to overwhelm the former, but not, I hope, today. We must all be grateful for the clear exposition of an important problem, for the practical value of the results obtained and for the possibilities opened up for future investigations. It is a real pleasure, therefore, for me to propose the vote of thanks today.

Dr J. Hartigan: I would like to suggest a non-parametric approach to Box and Cox's problem. Suppose in the ith experiment we observe y_i under conditions x_i and that it is desired to find the probability distribution of y given x for various x. The only general principle that seems to apply is a similarity principle—"What will happen under present circumstances will probably be similar to what happened under similar circumstances in the past" or more simply "like equals likely". The Meteorological Office does seem to be acting according to this principle in its long-range forecasts, where the procedure is to look at this month's weather, look in the records for a similar month, see what happened the following month then, and predict the same thing will happen next month, now—they would say, to predict what y_0 will be under conditions x_0, look among the (y_i, x_i) for an x_i close to x_0, then predict $y_0 = y_i$.

It does seem possible to offer a non-parametric method for predicting a new y at x_0; in least squares theory this would be the fitted value Y_0. The general procedure is to smooth from the various readings (y, x) in the neighbourhood of x_0, values of y being given greater or less weight according to x's "similarity" to x_0; just how the weights are to be chosen, or how the y's are to be combined is an open question; the least squares answer is $Y_0 = \Sigma\alpha_i y_i$, where the weights α_i (possibly negative, but not very, and nearly always adding to one) are calculated from the linear model.

Box and Cox are assuming that for some transformed set of observations $f(y_i)$, the model is valid, and their smoothed value would be given by

$$f(Y_0) = \Sigma\alpha_i f(y_i).$$

A "non-parametric" approach would be to order the observations $y_{(1)}, \ldots, y_{(n)}$ and select Y_0 such that

$$\sum_{y_{(i)} < Y_0} \alpha_i = \tfrac{1}{2}\Sigma\alpha_i.$$

Essentially, Y_0 is the median of the distribution consisting of points $y_{(i)}$ with probability α_i (possible negative values confuse this interpretation). The justification of this procedure is that Y_0 should not be too far from the value obtained by Box and Cox's procedure, since the median of the $f(y_i)$'s will be approximately equal to the mean of the $f(y_i)$'s; but this procedure is invariant under *any* monotonic transformation of the observations.

I have tried this with Box and Cox's 3^3 experiment, when x_0 is at the centre of the cube (0, 0, 0). The weights α_i will depend on the linear model; for a complete factorial model $\alpha_i = 1$ at (0, 0, 0) and 0 elsewhere so that no smoothing takes place; for the second-degree polynomial model $\alpha_i = 7$ at the centre, 4 at the midpoint of a face, 1 at the midpoint of an edge and -2 at a vertex; for the first- and zero-degree polynomials, $\alpha_i = 1$ everywhere and the smoothing is excessive.

The smoothed values with various similarity coefficients (we may regard α_i as the relevance of the ith observation to Y_0) and various methods of combination are

Degree of Polynomial	*Mean*	*Mean* log	*Median*
0,1	861	564	566
2	724	610	604
CF	620	620	620

Negative weights are a nuisance, and, also, we would like the similarity coefficients to decrease with distance. However, least squares is the only general way of generating the coefficients at present.

I wonder if the interquartile range of the distribution over the y_i with weights α_i would be a reasonable (transformation invariant) measure of dispersion of a new observation y about Y_0. In general this would tend to be large if y_i's which were observed under highly similar conditions were a long way from the predicted Y_0 at x_0.

A preliminary analysis of the above type based on the order statistics would be invariant under monotonic transformation, and so would seem an appropriate method of finding a transformation in which an ordinary "metric" analysis might be performed.

I have found this paper extremely informative and stimulating and it gives me great pleasure to second the vote of thanks to Professors Box and Cox.

The vote of thanks was put to the meeting and carried unanimously.

The following written contribution was read by Professor D. G. Kendall.

Professor J. W. Tukey: The results reported by Professors Box and Cox clearly represent a substantial step forward; all those concerned with the actual analysis of data should be pleased to know that they do exist, both because of the new and modified techniques which they urge us to try, and because these results were obtained by using almost "all the allowed principles of witchcraft" as of the year 1964: normality assumptions, maximum-likelihood estimation, Bayesian inference and *a priori* distributions invariant under natural, transitive groups. This last fact makes it inevitable that intelligent choice of modes of expression for the observed responses will become both socially acceptable and widely taught and that the long-run consequences for the analysis of data will be very desirable.

While this is a useful step forward, it is, I think, important not to overestimate its conclusiveness. From the point of view of the man who does indeed have data to analyse, these results are merely further guidance about a situation only reasonably close to the one he actually faces. This is, of course, no novelty in statistics, but some aspects of the present discussion make it important to re-emphasize some things that should be familiar to all of us. In the authors' discussion, as in all to nearly all of our presently available theory, all the approaches are at least formally based upon a model involving normality—or, as I would rather say, Gaussianity. I think that this is stressed by the discussion in Section 5 where one is asked to look first at the evidence from assumed Gaussianity, then at the evidence from an additional assumption of constancy of variance in the presence of Gaussianity and, finally, at the evidence from a further assumption of additivity in the presence of both other assumptions. So long as we are going to work with tight specifications, where only a few parameters can be allowed to enter, it is hard to see how things can be done in any other way than this. But from the point of view of the man with the actual data, it would make much more sense to ask—possibly in vain—for an analysis in which one could examine first the evidence derived from assumed additivity in the absence of other assumptions, secondly (in those situations where this was appropriate) the evidence provided by an additional assumption of constant variance in the presence of additivity, and thirdly (in perhaps a few cases) the additional evidence provided by assumed Gaussianity, in the presence of both additivity and constancy of variance. (If additivity —or, more generally, parsimony—is at issue, considerations of constancy of variance and Gaussianity of distribution are usually negligible, at least so far as the choice of a mode of expression is concerned. If additivity is not at issue, constancy of variance usually dominates Gaussianity of distribution.) If all of us can have enough good ideas over a long enough period of time, perhaps we can come, eventually, to a theory which corresponds more directly to what we desire. It may well be that, with the exception of very rare instances, the differences in practice associated with such an approach would be inappreciably different from those suggested by the present approach. The widespread tendency for additivity, constancy of variance and Gaussianity of distribution to come and go as a group offers us such a hope. It would be nice to know whether or not this hope is justified.

We are all used to having maximum-likelihood estimation combine different bits of evidence with quite appropriate weights. Accordingly, we may hope that this is still the case in the present situation, but I must report that the relative weighting of the evidence provided by interaction sums of squares and error sums of squares does not feel as if it were being quite fairly weighted when one merely looks, as in Table 3, at the total of these two sums of squares. Perhaps the decomposition into the three parts mentioned above, and concentration upon the part associated with the additivity assumption, might produce a much heavier weighting of the interaction sums of squares. Again it would be interesting to know whether or not this is true.

In most circumstances one is going to be more interested in reaching additivity than in maximizing the formal sensitivity of the main effects. There will be, however, a few instances where the reverse is true. I am not clear, from the discussion of Table 6, to what extent the results of applying the proposed approach rigorously and without thought will differ from the results obtained by seeking maximum sensitivity. If there should be differences which persist as the amount of data is increased without limit, I think one will have, in the long run, to look more carefully into the choice of criterion, where a decision to look need not imply an ultimate decision to adopt a different criterion.

Clearly Box and Cox have made a major step forward in the succession of approximations which give us better and better answers to an important problem of practice.

The following written contribution was read by the Honorary Secretary.

Professor R. L. PLACKETT: The authors have come up with the interesting ideas we would have expected from them, and deserve our congratulations for a paper which will be widely appreciated. They have made full use of modern computational facilities and the two systems of inference which are currently competing for our attention. An impression left by reading their paper is that the data should be fed into a large and powerful machine which will very quickly draw all the necessary graphs and print out the best analysis of variance available in the circumstances. Those accustomed to the blissful ease of the standard analysis of variance calculations will need to be convinced that such hard work is really necessary, and will ask for assurance that too much responsibility has not been delegated.

So much has recently been said on Bayesian procedures that it is a relief to find that the authors are not really Bayesians at all, but have been very ingenious in using Bayesian arguments without ever becoming fully committed to them. Thus they call for uniform distributions, but only over the region where the likelihood is appreciable, and they justify their preference for a Bayesian procedure on the grounds that the confidence coefficients from asymptotic distribution theory are closer to their nominal values if L_b is used instead of $L_{\max}$. It is true that in the further analysis separating out A and H they suggest that the two procedures may lead to appreciably different conclusions, but the circumstances in which this might occur are not closely defined. Surely it is not the magnitude of either $S_{\nu_r}(\lambda; z)$ or $F(\lambda; z)$ which is relevant, but that of the derivatives of these quantities with respect to λ. In any case, the authors do not tell us what they would do if the conclusions differ markedly; but it accords with the spirit of this long-awaited collaboration that we should be left in doubt as to which method of inference to follow.

Likelihood procedures have also been well publicized and discussed, but there is a practical point which seems not to have been emphasized in the midst of a good deal of mathematical and logical argument. It arises because the likelihood function contains much that is taken for granted in the way of distributional forms, and is no substitute for an inspection of the data. As a simple illustration, consider a large sample of measurements in which half are clustered round the value a and half round the value b $(a \neq b)$. The assumption that this constitutes a sample from a normal distribution with mean μ and standard deviation σ leads to an exactly parabolic log likelihood function for μ, but the inferences that this would suggest conflict with those obtained directly from the data.

It is tempting to contrast the smooth and deceptive character of a likelihood function with the spotty but straightforward nature of Anscombe and Tukey's procedures. They fit a full linear model to the original data and plot residuals against fitted values. Residuals are something which the authors have not calculated, but it would have been interesting to see other methods at work on the same examples. One might consider a modification of the Anscombe–Tukey procedure in which the predicted value Y is plotted against the observed value y. This will lead to a linearizing transformation $Y = f(y)$ (e.g. by Dolby's, 1963, analysis of the simple family); the procedure can be iterated if necessary and should converge under reasonable conditions. It may be objected that the possibility of differing variances is not taken into account, but the usual argument is that the same transformation does for both. If a greatly differing transformation is necessary to equalize the variances, then the experiment is unlikely to be very successful.

In the second part of their paper, the authors separate the contributions of linearity, constant variance and normality, but the place of normality in their analysis is logically different from that occupied by the other two, since normality is not a constraint which they either apply or relax. For that, they would presumably need to carry through the entire analysis with some other distribution.

Professor M. S. Bartlett: Like Professor Tukey, I think that the authors have made a major step forward in this paper on the theory of transformations. I think also, like Professor Plackett, I was a little uneasy about the extent to which complicated analysis might seem necessary.

Again, like Mr Nelder, I found myself wondering about the Box and Cox nature of the paper and in particular whether this kind of oscillatory character between likelihood and Bayes analysis had any relevance to the Box and Cox aspect! Perhaps Professor Cox may wish to comment on this; on this point of Bayes versus likelihood I would especially welcome his views on whether he is advocating them as equally useful or whether he has reached any conclusions as to whether one is better than the other. In particular I would certainly draw attention to the point made in the paper, and I think Professor Plackett made this point also, that whichever analysis you make, the inference is very conditional on your set of assumptions from which you start.

Now to come to other minor points, I think I have only two to make. One was in the approximation used for the log likelihood, the max log likelihood and the use of χ^2 with this, and I wondered whether Professor Cox, or for that matter, Professor Box, could make any comment on the accuracy in this in other than very large samples. One knows that the distribution is valid up to but not including order $1/n$, and one knows, for example, from Professor Box's work, that if you want to go to order $1/n$ you have to bring in a different multiplying factor to your χ^2 approximation. And it would help to know whether there is any possibility of getting the sort of confidence limits based on the χ^2 analysis a bit more exact, and if not, how misleading they might occasionally be.

I think my last point is one that was raised by Professor Tukey and that is, I did wonder about the uniqueness of this order of taking the various factors, normality, additivity and homogeneity of variances, and whether you would reach anything like the same sort of conclusion if you tried to take them in a different order.

Dr M. R. SAMPFORD: Like Professor Tukey, I am rather nervous about the effect of the assumed normality of the transformed variable on the additivity, in particular, and to a lesser extent on the homogeneity of variance, when in fact no single transformation will achieve all three properties. The relatively small amount of information about λ obtained from the normality assumption in the example (Table 8) seems to be reassuring on this point, but the possible effects when the transformed distribution is rather far from normal might still be serious. Of course, one can sometimes advance a more plausible distributional model, and in this context it may be worth suggesting that, though the title of this paper should more properly be "An Analysis of Transformations to Normality", the ingenious approach on which it is based could perfectly well be applied to other distributions. For example, I have several times encountered response-time distributions—in particular, distributions of time to death—that appear log-normal at the lower end of the scale, but have a secondary mode in the upper tail. This might suggest that some animals die as a direct result of damage caused by the treatment, but that others, having a high tolerance or being, by chance, little damaged, may survive the initial shock, only to die later as a result of physiological disturbance caused by the damage. One might, by making some assumptions about distributions of damage and tolerances, derive a more or less plausible class of distributions for transformed times that might be expected to be consistent with variance homogeneity and at least approximate additivity. The method of this paper could then be applied to determine the most satisfactory transformation leading to a distribution in this class. This is perhaps a rather extreme example, but I hope suggests the potential value of the authors' approach in situations where additivity need not be expected to involve, as it often does, near-normality.

Dr C. A. B. SMITH: I merely wish to draw attention to a recent paper by A. F. Naylor (1964). He applied the arcsine, logit, log-log and normal equivalent deviate transformations to four sets of biological data. He concluded that for all practical purposes they could be considered as equivalent. For example, in most of the entries the expected numbers calculated from the four transformations differ only slightly in the first decimal place.

Mr D. KERRIDGE: I have two comments to make, one general and one particular. The general comment is that it is very pleasant to have a paper in which the idea is obvious. I am not saying this in any derogatory sense. I think all the great ideas were obvious ones. Nothing could be more obvious than the idea of taking a parametric family and estimating the parameter. It is strange that such an obvious idea should take such a long time to be seen, but in many ways, the simpler the idea, the greater the discovery. There is, for example, much more chance that a simple idea will be used in practice. The particular comment concerns the rather strange prior distribution which has the interesting property that it contains the observations. We cannot let the night go without saying something about that. Clearly this is not an expression of belief, so some people would not call it a probability. It is not prior, because it is determined *a posteriori*, and so it is a pseudo-prior pseudo-probability. Now I am not against it because of its strangeness, since obviously the authors have extremely good reasons for using it. They use it because it works. It is very interesting indeed to find a practical example in which you have to use something which clearly is a pseudo-probability. I believe that as we get to use Bayes's theorem instead of talking about it, as I hope we are going to do in the future, we are going to come up against many more of these peculiar things. For example, I think that to get sensible significance tests in Bayesian theory we are going to have to use prior probabilities which depend on the number of observations. These again will be pseudo-probabilities, in a sense pseudo-prior too. So this is a very interesting first example of something which will eventually, I think, shed some light on what probabilities really are. My view is that they do not express beliefs. They are a convenient figment introduced to do something we do not really understand yet, but by examining examples of this sort I hope that one day we will achieve understanding.

Mr E. M. L. BEALE: I should like to add my thanks to Professors Box and Cox for a most valuable paper, and to ask one question. Would the authors ever consider using a transformation of the type (1) when some y's are negative, or one of type (2) where some $y_i + \lambda_2$ is negative? Such a transformation obviously has strange arithmetic properties. It gives a real answer if λ_1 is integral, and I think one can always overcome any problems created by the fact that y may not be uniquely determined by the value of $y^{(\lambda)}$. But would the transformation ever make sense statistically?

The following written contribution was received after the meeting:

Professor F. J. ANSCOMBE: The authors are to be congratulated on a most remarkable paper. The basic idea is highly original, and the tackling of horrendous difficulties is breath-taking. The examples are illuminating, and the preliminary "rather informal" analysis of the textile example is statistry in the grand manner—but, indeed, the whole paper is that.

Because of my own efforts with residuals, I have been particularly interested by Section 6. In my 1961 paper I gave a formula for roughly estimating the power transformation that would remove Tukey's type of removable non-additivity, and also one for estimating the power transformation that would remove an exponential dependence of error variance on the mean. The formulas were based essentially on the statistics denoted by T_{12} and T_{21}, respectively, in this paper. I did *not* also give a formula aimed at removing skewness of the error distribution, based on the statistic here denoted by T_{30}, though I have since used such a formula; in the notation of my 1961 paper the formula goes

$$p = 1 - 2g_1 \bar{y}/3(2+g_2)\, s.$$

(My p is Box and Cox's λ, $\bar{y}$ is the overall sample mean, s the residual root mean square, and g_1 and g_2 are analogues of Fisher's g-statistics.) It was my thought that one would calculate one or more of these expressions, and (if more than one) hope they would somewhat agree. No doubt, with factorial data showing pronounced effects for at least two factors, one would attach primary importance to additivity. With only one effective factor, there would be no question of additivity, and one would attach primary importance to constancy of variance. With no effective factors, and in particular with a simple homogeneous sample, there would be nothing to worry about except skewness.

Now Professors Box and Cox have shown that these three separate estimates should (very nearly) be averaged in a certain proportion to yield a best estimate of the power. This result, for the relatively simple calculations based on residuals from a least-squares analysis on one scale, parallels the subtle decomposition of the likelihood function into three parts in Section 5.

Professor Cox replied briefly at the meeting and the authors subsequently replied more fully in writing as follows:

We are very grateful to the speakers for their encouraging and helpful remarks.

One important general issue raised by Professors Tukey, Plackett, Bartlett and Dr Sampford concerns priorities for the criteria of simplicity of the model and specifically of additivity, A, homogeneity of variance, H, and normality, N. We certainly agree on the importance of the first of these, as indeed we indicate in our remarks at the end of Section 2. In the formal analysis of Section 5 we have considered N, HN, AHN as three models in that order. If one is to employ a parametric approach one must, it seems, start from some distributional assumption although, of course, if desired this could be broader than that adopted here. Furthermore, there is no reason in principle why A should not have been taken before H in discussing the biological example. We would then have to fit an additive model with separate within-cell variances. The rough justification for thinking that the procedure given in the paper genuinely separates out the effects of N, H and A is that $M(\lambda; \mathbf{z})$, on which (47) and (51) depend, is a valid descriptive measure of heterogeneity of variance independently of N. Likewise $F(\lambda; \mathbf{z})$ is a descriptive measure of non-additivity independently of H and N. If we started from a non-normal model, we would get a different measure of heterogeneity of variance, but except in extreme circumstances it is unlikely that it would be minimized by a value of λ very different from that minimizing $M(\lambda; \mathbf{z})$. An analogous remark applies to $F(\lambda; \mathbf{z})$. Under non-normality the weighting of the different requirements will be different, but it is hard to see how a radically different value of λ could emerge from the final analysis.

Concerning Professor Tukey's point about the appropriateness of the weighting given by the likelihood in the biological example, the truth seems to be that in this example non-additivity is not in fact the major contribution in determining λ. The sizes of the mean squares in Table 3 seem rather to bear this out than to contradict it. Concerning Tables 3 and 6 a striking thing is not only the removal of non-additivity, or correspondingly in Table 6 the simplification of the model, but also the large increase in sensitivity of the experiment. The result achieved by transformation is in fact equivalent to threefold increase in experimental effort.

In the paper we were at pains to stress that, where the procedures do seem relevant, we recommend using them in a flexible way, and that the assumptions on which they are based are a tentative working basis for the analysis rather than anything to be adopted irrevocably. In particular, in the discussion of the textile example we deliberately gave first the "common-sense" analysis before the more elaborate one. As Mr Kerridge has very rightly stressed, the basic idea is an extremely simple one; in particular, the absence of iterative calculations is a considerable practical advantage. We hope that this will reassure Professor Plackett that we are not advocating unnecessary elaboration. Mr Nelder has stated extremely clearly the need for a more searching examination of "assumptions".

We have not specifically investigated the point raised by Professor Bartlett concerning the adequacy of the chi-squared approximation for confidence intervals for λ. However, the line we have followed in finding a closer approximation to the posterior density of λ leads to posterior intervals based on the F distribution and a similar approximation might be found for confidence intervals. The use of $L_b(\lambda)$ instead of $L_{\max}(\lambda)$ was suggested by analogy with Bartlett's (1937) procedure of applying the likelihood-ratio procedure after suitable contrasts have been removed by transformation. The difficulty when λ is unknown is that the transformations to remove the parameters θ depend on λ, so that the argument is at best approximate. We were most interested in Mr Nelder's remarks on this point and hope that he will develop his ideas further.

The maximum-likelihood approach and the Bayesian approach have deliberately been given as entirely separate but parallel developments. Professor Plackett suggests that we justify the Bayesian approach only because it leads to "better" confidence intervals; this is not so. Several speakers have commented on the special prior distribution (19) which involves the observations. As we remarked in the paper, it is possible that there is an alternative and better approach to this; one way may be to make the prior distributions for the contrasts depend on the general population mean. However, the observations enter (19) only in a mild way in establishing the overall level of the observations, usually the overall geometric mean in our special cases. It is essential that some allowance should be made for the fact that the prior distribution for the magnitude of the contrasts depends on the overall magnitude of the observations.

In answer to Mr Beale's question, we feel that, while it is probably possible to develop the theory for non-monotonic transformations of the dependent variable, we cannot think of any situations where such transformations would be physically allowable.

We are grateful to Dr Smith for his reference to Naylor's work. However, Naylor seems to be considering situations where the transformations are, over the relevant range, practically linear functions of one another. In our examples the relative range of variation of the observations is high, the transformations are very non-linear and this is of course why we are able to obtain fairly sharp discrimination between the different values of λ. In the quantal response case, the transformations in question become essentially different only in the tails of the response curve, and observations there would be required for the differences to be detectable and of practical importance.

We are very interested in Professor Anscombe's remarks on residuals. Further comparisons of the analysis of residuals with the methods of our paper would be of value.

We are interested in Dr Hartigan's problem and formulation. However, this seems essentially different from ours, partly because in our applications we are primarily interested in changes in response, rather than in absolute responses, and partly because one of our primary objectives is to find a scale on which the factor effects are succinctly characterized by a few parameters. Even if the distributional assumptions were to be phrased non-parametrically (which we would in any case not wish to do), we must have parameters in order to describe at all concisely the changes in response in a complex system.

References in the Discussion

Dolby, J. L. (1963), "A quick method for choosing a transformation", *Technometrics*, **5**, 317–326.

Naylor, A. F. (1964), "Comparisons of regression constants fitted by maximum likelihood to four common transformations of binomial data", *Ann. hum. Genet., Lond.*, **27**, 241–246.