EXAMINATION OF RESIDUALS

F. J. ANSCOMBE

PRINCETON UNIVERSITY AND THE UNIVERSITY OF CHICAGO

1. Introduction

1.1. Suppose that n given observations, $y_1, y_2, \cdots, y_n$, are claimed to be independent determinations, having equal weight, of means $\mu_1, \mu_2, \cdots, \mu_n$, such that

$$\mu_i = \sum_r a_{ir}\theta_r, \tag{1}$$

where $\mathbf{A} = (a_{ir})$ is a matrix of given coefficients and (θ_r) is a vector of unknown parameters. In this paper the suffix i (and later the suffixes j, k, l) will always run over the values $1, 2, \cdots, n$, and the suffix r will run from 1 up to the number of parameters (θ_r).

Let $(\hat{\theta}_r)$ denote estimates of (θ_r) obtained by the method of least squares, let (Y_i) denote the fitted values,

$$Y_i = \sum_r a_{ir}\hat{\theta}_r, \tag{2}$$

and let (z_i) denote the residuals,

$$z_i = y_i - Y_i. \tag{3}$$

If A stands for the linear space spanned by $(a_{i1}), (a_{i2}), \cdots$, that is, by the columns of $\mathbf{A}$, and if $\overline{A}$ is the complement of A, consisting of all n-component vectors orthogonal to A, then (Y_i) is the projection of (y_i) on A and (z_i) is the projection of (y_i) on $\overline{A}$. Let $\mathbf{Q} = (q_{ij})$ be the idempotent positive-semidefinite symmetric matrix taking (y_i) into (z_i), that is,

$$z_i = \sum_j q_{ij}y_j. \tag{4}$$

If A has dimension $n - \nu$ (where $\nu > 0$), $\overline{A}$ is of dimension ν and $\mathbf{Q}$ has rank ν. Given A, we can choose a parameter set (θ_r), where $r = 1, 2, \cdots, n - \nu$, such that the columns of $\mathbf{A}$ are linearly independent, and then if $\mathbf{V}^{-1} = \mathbf{A}'\mathbf{A}$ and if $\mathbf{I}$ stands for the $n \times n$ identity matrix (δ_{ij}), we have

$$\mathbf{Q} = \mathbf{I} - \mathbf{AVA}'. \tag{5}$$

The trace of $\mathbf{Q}$ is

$$\sum_i q_{ii} = \nu. \tag{6}$$

Research carried out at Princeton University and in part at the Department of Statistics, University of Chicago, under sponsorship of the Logistics and Mathematical Statistics Branch, Office of Naval Research.

The least-squares method of estimating the parameters (θ_r) is unquestionably satisfactory under the following

Ideal statistical conditions. The (y_i) are realizations of independent chance variables, such that the $(y_i - \mu_i)$ have a common normal distribution with zero mean.

That is to say, given the ideal conditions, the least-squares estimates of the parameters, together with the residual sum of squares, constitute a set of sufficient statistics, and all statements or decisions resulting from the analysis will properly depend on them. They may also depend on a prior probability or loss system, and estimates eventually quoted may therefore differ from the least-squares estimates, as C. M. Stein has shown.

We shall refer to the differences $(y_i - \mu_i)$ by the conventional name of "errors," and denote the variance of the error distribution by σ^2. Under the ideal conditions, $\mathbf{Q}\sigma^2$ is the variance matrix of the residuals (z_i), which have a multivariate normal chance distribution over $\bar{A}$. We shall denote by s^2 the residual mean square, $(\sum_i z_i^2)/\nu$, which under the ideal conditions is an unbiased estimate of σ^2.

1.2. The object of this paper is to present some methods of examining the observed residuals (z_i), in order to obtain information on how close the ideal conditions come to being satisfied. We shall first (section 2) consider the residuals in aggregate, that is, their empirical distribution, and then (section 3) consider the dependence of the residuals on the fitted values, that is, the regression relations of the pairs (Y_i, z_i). In each connection we shall propose two statistics, which can be taken as measures of departure from the ideal conditions and can be used as criteria for conventional-type significance tests, the ideal conditions being regarded as a composite null hypothesis. Section 4 will be concerned with justifying the statistics proposed, and section 5 with examples of their use.

The problem of examining how closely the ideal conditions are satisfied is very broad. Despite the immense use of the least-squares methods for well over a century, the problem has received no comprehensive treatment. Particular aspects have been considered by many authors, usually on a practical rather than theoretical level. This paper, too, is concerned with particular aspects, except for a little general discussion in section 6. The reader will appreciate that it is not always appropriate to base an investigation of departures from the ideal conditions on an examination of residuals. This will be illustrated below (section 5.5). And if residuals are examined, many other types of examination are possible besides those presented here. For example, when the observations have been made serially in time, and time is not among the parameters (θ_r), interesting information may possibly be obtained by plotting the residuals against time, as Terry [17] has pointed out. There are circumstances in which the error variance may be expected to be different for different levels of some factor, as for example in a plant-breeding experiment when the lines compared differ markedly in genetic purity. The ways in which the ideal conditions can fail to obtain are of course countless.

The material of this paper has been developed from two sources: first, some

unpublished work relating primarily to the analysis of data in a row-column cross-classification, done jointly with John W. Tukey [1], [4]—to him is due the idea of considering simple functions of the residuals, of the type here presented, as test criteria; second, a study of correlations between residuals, in connection with an investigation of rejection rules for outliers [3]. Familiarity with the latter is not assumed, but overlap has been kept to a minimum, with the thought that a reader interested in this paper will read [3] also.

1.3. The methods developed below appear not to have any sweeping optimum properties. They are easiest to apply, and possibly more particularly appropriate, if the following two conditions are satisfied.

Design condition 1. *A contains the unit vector, or (in other words) the parameter set* (θ_r) *can be so chosen that one parameter is a general mean and the corresponding column of* **A** *consists entirely of* 1's.

Design condition 2. *The diagonal elements of* **Q** *are all equal, thus* $q_{ii} = \nu/n$ *(for all* i*).*

These are labeled "design conditions" because they are conditions on A. If the observations come from a factorial experiment, A depends on the design of the experiment, in the usual sense, and also on the choice of effects (interactions, and so forth) to be estimated, so the name is not entirely happy, but will be used nevertheless.

Condition 1 is, one supposes, almost always satisfied in practice. If it were not, and if the residuals were to be examined, the first idea that would occur would be to examine the mean of the residuals, and that would be equivalent to introducing the general mean as an extra parameter, after which condition 1 would be satisfied. So this condition is not much of a restriction. A consequence of the condition is that every row (or column) of **Q** sums to zero, that is, $\sum_j q_{ij} = 0$ for each i. Hence if $\bar{\rho}$ denotes the average correlation coefficient between all pairs (z_i, z_j), where $i \neq j$, we have

$$1 + (n - 1)\bar{\rho} = 0. \tag{7}$$

The residuals themselves sum to zero, and so their average $\bar{z} = 0$.

Condition 2 is satisfied for a broad class of factorial and other randomized experimental designs, provided that the effects estimated do not depend for their definition on any quantitative relation between the levels of the factors. In other circumstances we shall expect condition 2 not to be satisfied exactly. A consequence of condition 2 and the idempotency of **Q** is that the sum of squares of entries in any row (or column) of **Q** is the same, namely

$$\sum_j (q_{ij})^2 = q_{ii} = \frac{\nu}{n}. \tag{8}$$

Hence if $\overline{\rho^2}$ is the average squared correlation coefficient between all pairs (z_i, z_j), where $i \neq j$, we have

$$1 + (n - 1)\overline{\rho^2} = \frac{n}{\nu}. \tag{9}$$

Condition 2 was imposed in the study of outliers [3], in order to avoid any question as to the correct weighting of the residuals. Ferguson [10] has considered outliers when condition 2 is not satisfied.

In the next two sections we shall proceed first without reference to conditions 1 and 2, and then we shall see how the expressions obtained reduce when conditions 1 and 2 are introduced.

2. Empirical distribution of residuals

2.1. *Skewness.* Let us suppose that the ideal statistical conditions obtain, as in the above enunciation, except that we delete the word "normal." Let γ_1 and γ_2 be the first two scale-invariant shape coefficients (supposed finite) of the error distribution, measuring skewness and kurtosis, defined as

$$\gamma_1 = \frac{1}{\sigma^3} E[(y_i - \mu_i)^3], \qquad \gamma_2 = \frac{1}{\sigma^4} \{E[(y_i - \mu_i)^4] - 3\sigma^4\}. \tag{10}$$

We shall study estimates g_1 and g_2 of γ_1 and γ_2, based on the statistics s^2, $\sum_i z_i^3$, $\sum_i z_i^4$, analogous to Fisher's statistics [13] for a simple homogeneous sample.

Since we suppose the errors $(y_i - \mu_i)$ to be independent with zero means, we have at once

$$E(\sum_i z_i^3) = \sum_{ij} (q_{ij})^3 \gamma_1 \sigma^3. \tag{11}$$

Provided $\sum_{ij}(q_{ij})^3 \neq 0$, we can define g_1 by

$$g_1 = \frac{\sum_i z_i^3}{\sum_{ij} (q_{ij})^3 s^3}. \tag{12}$$

We now consider the sampling distribution of g_1 under the full ideal conditions, so that $\gamma_1 = 0$. The distribution of (z_i) has spherical symmetry in $\overline{A}$, and the radius vector s is independent of the direction. Hence g_1, being a homogeneous function of (z_i) of degree zero, is independent of s, and we have

$$E(g_1) = 0, \qquad \text{Var } (g_1) = \frac{E[(\sum_i z_i^3)^2]}{[\sum_{ij} (q_{ij})^3]^2 E(s^6)}. \tag{13}$$

It is well known that $E(s^6) = (\nu + 2)(\nu + 4)\sigma^6/\nu^2$. As for the numerator, we have $E[(\sum_i z_i^3)^2] = \sum_{ij} E(z_i^3 z_j^3)$. Now, whether i and j are the same or different, z_i and z_j have a joint normal distribution with variance matrix

$$\sigma^2 \begin{pmatrix} q_{ii} & q_{ij} \\ q_{ij} & q_{jj} \end{pmatrix}. \tag{14}$$

It follows that $E(z_i^3 z_j^3)$ is the coefficient of $t_1^3 t_2^3/(3!)^2$ in the expansion of the moment-generating function $\exp[(1/2)\sigma^2(q_{ii}t_1^2 + 2q_{ij}t_1t_2 + q_{jj}t_2^2)]$. We obtain easily

$$E(z_i^3 z_j^3) = \{6(q_{ij})^3 + 9q_{ij}q_{ii}q_{jj}\}\sigma^6. \tag{15}$$

Hence

$$\text{Var}\,(g_1) = \frac{[6\sum_{ij}(q_{ij})^3 + 9\sum_{ij} q_{ij}q_{ii}q_{jj}]\nu^2}{[\sum_{ij}(q_{ij})^3]^2(\nu+2)(\nu+4)}. \tag{16}$$

Because $\mathbf{Q}$ is positive-semidefinite, the expression $\sum_{ij}q_{ij}q_{ii}q_{jj}$ is nonnegative. It vanishes under design conditions 1 and 2, or (more generally) if the vector (q_{ii}) lies in A. Then (16) reduces to

$$\text{Var}\,(g_1) = \frac{6\nu^2}{(\nu+2)(\nu+4)\sum_{ij}(q_{ij})^3}. \tag{17}$$

If $\overline{\rho^3}$ denotes the average cubed correlation coefficient between pairs (z_i, z_j), where $i \neq j$, (17) can be expressed under condition 2 as

$$\text{Var}\,(g_1) = \frac{6n^2}{\nu(\nu+2)(\nu+4)\{1+(n-1)\overline{\rho^3}\}}. \tag{18}$$

Under condition 2 g_1 itself can be expressed as

$$g_1 = \frac{n^2\sum_i z_i^3}{\{1+(n-1)\overline{\rho^3}\}(\nu\sum_i z_i^2)^{3/2}}. \tag{19}$$

For the simple homogeneous sample, $\nu = n-1$ and $\overline{\rho^3} = -1/\nu^3$, and (18) reduces to Fisher's result,

$$\text{Var}\,(g_1) = \frac{6n(n-1)}{(n-2)(n+1)(n+3)}. \tag{20}$$

For a row-column cross-classification with k rows and l columns, $n = kl$ and $\nu = (k-1)(l-1)$, and we find

$$1 + (n-1)\overline{\rho^3} = \frac{n(k-2)(l-2)}{\nu^2}. \tag{21}$$

Hence (18) gives, provided k and l both exceed 2,

$$\text{Var}\,(g_1) = \frac{6n\nu}{(\nu+2)(\nu+4)(k-2)(l-2)}. \tag{22}$$

If n and ν are both large, it is commonly (but not invariably) the case that $1 + (n-1)\overline{\rho^3}$ is very close to 1, and then the right side of (18) is roughly $6n^2/\nu^3$, about the same as the variance of g_1 for a homogeneous sample of size $\nu(\nu/n)^2$.

In principle it is possible by the same method to find higher moments of the sampling distribution of g_1, under the full ideal conditions. It is easy to see that the odd moments vanish. The fourth moment is as follows.

$$E(g_1^4) = \frac{108\nu^5}{(\nu+2)(\nu+4)(\nu+6)(\nu+8)(\nu+10)[\sum_{ij}(q_{ij})^3]^4}\Big\{[\sum_{ij}(q_{ij})^3]^2 + 18\sum_{ijkl}(q_{ij})^2(q_{kl})^2q_{ik}q_{jl} + 12\sum_{ijkl}q_{ij}q_{ik}q_{il}q_{jk}q_{jl}q_{kl} + 36\sum_{ijkl}q_{ij}q_{jk}q_{jl}(q_{kl})^2q_{ii} + 18\sum_{ijkl}q_{ik}q_{jl}(q_{kl})^2q_{ii}q_{jj} + 6\sum_{ijkl}q_{il}q_{jl}q_{kl}q_{ii}q_{jj}q_{kk} + 3\sum_{ij}q_{ij}q_{ii}q_{jj}\sum_{kl}(q_{kl})^3 + \frac{9}{4}(\sum_{ij}q_{ij}q_{ii}q_{jj})^2\Big\}. \tag{23}$$

Under conditions 1 and 2 the last five of the eight terms inside the braces vanish, leaving only the first three. Unfortunately the second and third terms are formidable to evaluate, in general. By way of considering a relatively simple special case, let us impose the further design condition that all the off-diagonal elements of **Q** are equal, barring sign. (Such designs are illustrated below in section 5.4; they include the simple homogeneous sample.) Then if we write c for ν/n, the elements in every row of **Q** consist of c in the diagonal and $\pm[c(1-c)/(n-1)]^{1/2}$ everywhere else, the number of minus signs exceeding the number of plus signs (in these off-diagonal elements) by $[c(n-1)/(1-c)]^{1/2}$, which must of course be an integer. Writing C for $c^2 - c(1-c)/(n-1)$, we find easily that

$$\sum_{ij}(q_{ij})^3 = \nu C, \tag{24}$$

and because $\sum_l (q_{kl})^2 q_{jl} = Cq_{jk}$ we have also

$$\sum_{ijkl}(q_{ij})^2(q_{kl})^2q_{ik}q_{jl} = C\sum_{ij}(q_{ij})^3 = \nu C^2. \tag{25}$$

I have been unable to evaluate completely the third term, $\Sigma_{ijkl}q_{ij}q_{ik}q_{il}q_{jk}q_{jl}q_{kl}$, except for the simple sample, having $\nu = n-1$ and $C = (n-2)/n$, when it can easily be shown to be $(n-1)(n-2)(n-3)/n^2$. On substitution into (23) we then verify Fisher's formula for $E(g_1^4)$ ([12], p. 22). For other designs having equal-magnitude correlations, we can say that

$$\sum_{ijkl}q_{ij}q_{ik}q_{il}q_{jk}q_{jl}q_{kl} = \nu c^5 - \frac{4\nu c^4(1-c)}{n-1} + \frac{3\nu c^3(1-c)^2}{n-1} - \frac{6\nu(n-2)c^3(1-c)^2}{(n-1)^2} + R, \tag{26}$$

where the successive terms on the right side are the contributions to the total sum from sets of suffixes (i, j, k, l) that are (respectively) all equal, all but one equal, equal in two different pairs, different except for one pair, and finally R stands for the balance from sets of wholly unequal suffixes. R consists of the sum of $n(n-1)(n-2)(n-3)$ terms each equal to $\pm[c(1-c)/(n-1)]^3$. Now if n is large and c not very close to 0 or 1, positive and negative values are roughly equally frequent among the elements of **Q**, and it seems highly plausible

that the terms of R almost cancel each other, so that $R = o(n)$. Assuming this, we find, for n large and c constant,

$$E(g_1^4) \sim \frac{108}{\nu^2 C^2}\left\{1 - \frac{12(1-c)}{\nu}\right\}, \qquad E(g_1^2) \sim \frac{6}{\nu C}\left\{1 - \frac{6}{\nu}\right\}. \tag{27}$$

Hence the kurtosis coefficient of the distribution of g_1 is asymptotically $36/n$, the same as for the simple sample of size n. Thus although the variance of g_1 exceeds that for a simple sample of the same size n by a factor of $(n/\nu)^3$ roughly, the shape of the distribution is apparently roughly the same. It is tempting to surmise that the similarity of shape may hold true fairly generally for complex designs such that all the off-diagonal elements of $\mathbf{Q}$ are small.

2.2. *Kurtosis.* Consider now the estimation of the kurtosis coefficient γ_2 of the error distribution, as posed at the outset of section 2. We find easily

$$E(\sum_i z_i^4) = \sum_{ij} (q_{ij})^4 \gamma_2 \sigma^4 + 3 \sum_i (q_{ii})^2 \sigma^4, \tag{28}$$

$$\nu^2 E(s^4) = E[(\sum_i z_i^2)^2] = \sum_i (q_{ii})^2 \gamma_2 \sigma^4 + \nu(\nu+2)\sigma^4. \tag{29}$$

We can therefore define g_2 by the following expression, provided the divisor D does not vanish,

$$g_2 = \left(\frac{\sum_i z_i^4}{s^4} - \frac{3\nu \sum_i (q_{ii})^2}{\nu+2}\right) D^{-1}, \tag{30}$$

where

$$D = \sum_{ij} (q_{ij})^4 - \frac{3[\sum_i (q_{ii})^2]^2}{\nu(\nu+2)}. \tag{31}$$

Under the full ideal statistical conditions, we have $E(g_2) = 0$ and

$$D^2 \operatorname{Var}(g_2) + \left(\frac{3\nu \sum_i (q_{ii})^2}{\nu+2}\right)^2 = \frac{E[(\sum_i z_i^4)^2]}{E(s^8)}. \tag{32}$$

Proceeding as before we find

$$E[(\sum_i z_i^4)^2] = \sum_{ij} E(z_i^4 z_j^4) = \{24 \sum_{ij} (q_{ij})^4 + 72 \sum_{ij} (q_{ij})^2 q_{ii} q_{jj} + 9[\sum_i (q_{ii})^2]^2\}\sigma^8 \tag{33}$$

and hence

$$\operatorname{Var}(g_2) = \frac{(24D + 72F)\nu^3}{D^2(\nu+2)(\nu+4)(\nu+6)}, \tag{34}$$

where

$$F = \sum_{ij} (q_{ij})^2 q_{ii} q_{jj} - \nu^{-1}[\sum_i (q_{ii})^2]^2. \tag{35}$$

Under condition 2, F vanishes and we have

$$g_2 = \frac{n^3}{\nu(\nu+2)\{1 + (n-1)\overline{\rho^4}\} - 3n}\left[\frac{\nu+2}{\nu}\frac{\sum_i z_i^4}{(\sum_i z_i^2)^2} - \frac{3}{n}\right], \tag{36}$$

$$\text{Var}\,(g_2) = \frac{24n^3}{[\nu(\nu+2)\{1+(n-1)\overline{\rho^4}\}-3n](\nu+4)(\nu+6)}. \tag{37}$$

For the simple sample we obtain Fisher's result,

$$\text{Var}\,(g_2) = \frac{24n(n-1)^2}{(n-3)(n-2)(n+3)(n+5)}. \tag{38}$$

For the cross-classification with k rows and l columns we have

$$\text{Var}\,(g_2) = \frac{24n^2\nu^2}{[(\nu+2)(k^2-3k+3)(l^2-3l+3)-3\nu^2](\nu+4)(\nu+6)}. \tag{39}$$

If n and ν are both large and $1+(n-1)\overline{\rho^4}$ is close to 1, the right side of (37) is roughly $24n^3/\nu^4$, about the same as the variance of g_2 for a simple sample of size $\nu(\nu/n)^3$.

It is possible to write down a general expression for $E(g_2^3)$, under the ideal conditions. We quote here only the reduced form under condition 2.

$$E(g_2^3) = \frac{1728\nu^5\left\{\sum_{ijk}(q_{ij})^2(q_{ik})^2(q_{jk})^2 - \frac{\nu^4(\nu+8)}{n^3(\nu+2)^2} - \frac{6\nu}{n(\nu+2)}D\right\}}{(\nu+2)(\nu+4)(\nu+6)(\nu+8)(\nu+10)D^3}. \tag{40}$$

For a design with equal-magnitude correlations, we find easily

$$\begin{aligned}\sum_{ijk}(q_{ij})^2(q_{ik})^2(q_{jk})^2 &= nc^6 + \frac{3nc^4(1-c)^2}{n-1} + \frac{n(n-2)c^3(1-c)^3}{(n-1)^2},\\ D &= nc^4 + \frac{nc^2(1-c)^2}{n-1} - \frac{3\nu c^2}{\nu+2}.\end{aligned} \tag{41}$$

When n is large, that is, as $n \to \infty$, with $c >$ a positive bound, these give

$$E(g_2^3) \sim \frac{1728}{nc^6}, \qquad \text{Var}\,(g_2) \sim \frac{24}{nc^4}, \tag{42}$$

and the skewness coefficient of the distribution of g_2 is asymptotically $6(6/n)^{1/2}$, the same as for a simple sample of size n.

3. Relation of residuals with fitted values

3.1. *Heteroscedasticity.* Let us suppose that a weakened form of the ideal statistical conditions obtains, namely: the (y_i) are realizations of independent chance variables, such that $y_i - \mu_i$ has a normal distribution with zero mean and variance proportional to $\exp(\chi\mu_i)$, where χ is a constant. Denoting the variance of y_i by σ_i^2, we suppose that $\sigma_i^2 \propto \exp(\chi\mu_i)$, so that $E(y_i - \mu_i)^2$ has a regression on μ_i. If χ is small, the regression is nearly linear. We shall study an estimate h of χ, for χ assumed small, based on the statistics s^2 and $\sum z_i^2 Y_i$, that is, on an empirical linear regression of (z_i^2) on (Y_i).

Let $\overline{Y} = \nu^{-1}\sum_i q_{ii}Y_i$, and let (r_{ij}) be the matrix taking (y_i) into $(Y_i - \overline{Y})$, that is,

$$(43) \qquad Y_i - \bar{Y} = \sum_j r_{ij} y_j, \qquad r_{ij} = \delta_{ij} - q_{ij} - \nu^{-1}(q_{jj} - \sum_k q_{jk} q_{kk}).$$

It is easy to show that $\sum_j r_{ij} q_{jk} = 0$ for all i and k. Under the assumed conditions, z_i and $Y_i - \bar{Y}$ (for any i) have a bivariate normal distribution, and

(44)

$$E(z_i) = 0, \qquad E[z_i(Y_i - \bar{Y})] = \sum_j q_{ij} r_{ij} \sigma_j^2, \qquad \text{Var}\,(Y_i - \bar{Y}) = \sum_j (r_{ij})^2 \sigma_j^2.$$

Now suppose that χ is small, so that

$$(45) \qquad \sigma_i^2 = \sigma^2[1 + \chi(\mu_i - \bar{\mu}) + O(\chi^2)],$$

where $\bar{\mu} = \nu^{-1} \sum_i q_{ii} \mu_i$. Then

$$(46) \qquad E[z_i(Y_i - \bar{Y})] = \sum_j q_{ij} r_{ij} (\mu_j - \bar{\mu}) \chi \sigma^2 + O(\chi^2).$$

Thus (except in the degenerate case when $r_{ij} = 0$ for all j) the regression coefficient of z_i on $Y_i - \bar{Y}$ is $O(\chi)$. Hence the (quadratic) regression coefficient of z_i^2 on $Y_i - \bar{Y}$ is $O(\chi^2)$, that is,

$$(47) \qquad E(z_i^2 | Y_i - \bar{Y}) = E(z_i^2) + O(\chi^2).$$

It follows that

$$\begin{aligned} (48) \qquad E[\sum_i z_i^2 (Y_i - \bar{Y})] &= \sum_i E(z_i^2) E(Y_i - \bar{Y}) + O(\chi^2) \\ &= \sum_{ij} (q_{ij})^2 \sigma_j^2 (\mu_i - \bar{\mu}) + O(\chi^2) \\ &= \sum_i \{q_{ii}\sigma^2 + \sum_j (q_{ij})^2 (\mu_j - \bar{\mu}) \chi \sigma^2\} (\mu_i - \bar{\mu}) + O(\chi^2) \\ &= \sum_{ij} (q_{ij})^2 (\mu_i - \bar{\mu})(\mu_j - \bar{\mu}) \chi \sigma^2 + O(\chi^2). \end{aligned}$$

This result suggests the estimate h of χ,

$$(49) \qquad h = \frac{\sum_i z_i^2 (Y_i - \bar{Y})}{\sum_{ij} (q_{ij})^2 (Y_i - \bar{Y})(Y_j - \bar{Y}) s^2}.$$

Naturally, if the matrix $((q_{ij})^2)$ is such that the denominator of (49) vanishes identically, or if all the (Y_i) are equal, we must consider that h does not exist. Failing that, the estimate h has a large-sample bias towards 0, since the denominator tends to be too large. In fact, when χ is small, s^2 is almost independent of the rest of the denominator, and we have

$$\begin{aligned} (50) \qquad E[\sum_{ij} (q_{ij})^2 (Y_i - \bar{Y})(Y_j - \bar{Y}) s^2] &= \sum_{ij} (q_{ij})^2 (\mu_i - \bar{\mu})(\mu_j - \bar{\mu}) \sigma^2 \\ &\quad + \sum_{ij} (q_{ij})^2 \,\text{Cov}\,(Y_i - \bar{Y}, Y_j - \bar{Y}) \sigma^2 + O(\chi^2), \end{aligned}$$

and

$$\begin{aligned} (51) \qquad \sum_{ij} (q_{ij})^2 \,\text{Cov}\,(Y_i - \bar{Y}, Y_j - \bar{Y}) &= \sum_{ijk} (q_{ij})^2 r_{ik} r_{jk} \sigma^2 + O(\chi) \\ &= \left[\frac{\nu - 1}{\nu} \sum_i (q_{ii})^2 - \sum_{ij} (q_{ij})^3 + \frac{1}{\nu} \sum_{ij} q_{ij} q_{ii} q_{jj}\right] \sigma^2 + O(\chi), \end{aligned}$$

as we see after some reduction. Thus the following might be regarded as a more satisfactory estimate of χ:

$$h^* = \frac{\sum_i z_i^2(Y_i - \bar{Y})}{\left[\sum_{ij} (q_{ij})^2(Y_i - \bar{Y})(Y_j - \bar{Y}) - \frac{\nu}{\nu + 2}\sum_{ijk} (q_{ij})^2 r_{ik} r_{jk} s^2\right] s^2}. \tag{52}$$

Actually we shall consider h rather than h^*, because it is simpler. The difference between h and h^* is likely to be negligible whenever there is enough dispersion among the (Y_i) to permit good estimation of χ.

We now consider the sampling distribution of h under the full ideal conditions, so that $\chi = 0$. The (z_i) and (Y_i) are completely independent. Let us consider the conditional distribution of h, given (Y_i). We find $E[h|(Y_i)] = 0$ and

$$\begin{aligned} \text{Var}\,[h|(Y_i)] &= \frac{\sum_{ij} E[z_i^2 z_j^2(Y_i - \bar{Y})(Y_j - \bar{Y})|(Y_i)]}{[\sum_{ij} (q_{ij})^2(Y_i - \bar{Y})(Y_j - \bar{Y})]^2 E(s^4)} \\ &= \frac{2\nu}{(\nu + 2)\sum_{ij} (q_{ij})^2(Y_i - \bar{Y})(Y_j - \bar{Y})}, \end{aligned} \tag{53}$$

$$E[h^3|(Y_i)] = \frac{8\nu^2 \sum_{ijk} q_{ij} q_{ik} q_{jk}(Y_i - \bar{Y})(Y_j - \bar{Y})(Y_k - \bar{Y})}{(\nu + 2)(\nu + 4)[\sum_{ij} (q_{ij})^2(Y_i - \bar{Y})(Y_j - \bar{Y})]^3}. \tag{54}$$

Let us see how these results simplify under special design conditions. Under conditions 1 and 2 we have $\bar{Y} = \bar{y}$, the simple average of the observations; and apropos of h^*, we have

$$\sum_{ijk} (q_{ij})^2 r_{ik} r_{jk} = \nu\left\{\frac{\nu - 1}{n} - \left(\frac{\nu}{n}\right)^2 [1 + (n - 1)\bar{\rho}^3]\right\}, \tag{55}$$

or roughly $\nu c(1 - c)$. Usually $\sum_{ij}(q_{ij})^2(Y_i - \bar{y})(Y_j - \bar{y})$ can be expressed in terms of the sums of squares appearing in an analysis of variance table, for which we use the notation SS( ). For example, for data in a one-way classification, l sets of k observations each, with a separate mean estimated for each set, so that $n = kl$, $\nu = (k - 1)l$, we find

$$\sum_{ij} (q_{ij})^2(Y_i - \bar{y})(Y_j - \bar{y}) = \frac{k - 1}{k}\,\text{SS(means)}, \tag{56}$$

where SS(means) stands for the sum of squares for differences between means, namely $k\sum_r(\bar{y}_r - \bar{y})^2$, where $\bar{y}_r$ is the mean of the rth set. For a cross-classification having k rows and l columns, $n = kl$, $\nu = (k - 1)(l - 1)$, we find

$$\begin{aligned} &\sum_{ij} (q_{ij})^2(Y_i - \bar{y})(Y_j - \bar{y}) \\ &\qquad = \frac{(k - 2)(l - 1)}{n}\,\text{SS(rows)} + \frac{(k - 1)(l - 2)}{n}\,\text{SS(columns)}. \end{aligned} \tag{57}$$

For a $k \times k$ Latin square, with $n = k^2$, $\nu = (k - 1)(k - 2)$, we find

$$\sum_{ij} (q_{ij})^2(Y_i - \bar{y})(Y_j - \bar{y}) = \frac{(k-2)(k-3)}{n} \text{SS(rows, columns, and letters).} \tag{58}$$

For a design having equal-magnitude correlations between the residuals, we find

$$\sum_{ij} (q_{ij})^2(Y_i - \bar{y})(Y_j - \bar{y}) = \frac{\nu(\nu-1)}{n(n-1)} \sum_i (Y_i - \bar{y})^2. \tag{59}$$

If n and ν are large and the correlations are close to being equal in magnitude, we may expect that as ν decreases with n fixed Var $[h|(Y_i)]$ will increase roughly in proportion to ν^{-2}, provided $\sum_i (Y_i - \bar{y})^2$ stays nearly constant—a more modest rate of increase than those of Var (g_1) and Var (g_2).

As for the third moment of h, the right side of (54) is an odd function of (Y_i), and is some measure of asymmetry of the set of fitted values. For a design satisfying conditions 1 and 2 and having equal-magnitude correlations, we find

$$\sum_{ijk} q_{ij}q_{ik}q_{jk}(Y_i - \bar{y})(Y_j - \bar{y})(Y_k - \bar{y}) = \left[c^3 - \frac{3c^2(1-c)}{n-1}\right] \sum_i (Y_i - \bar{y})^3 + R, \tag{60}$$

where R is the contribution to the sum from wholly unequal suffixes (i, j, k), consisting of $n(n-1)(n-2)$ terms each of which has the following form: $\pm[c(1-c)/(n-1)]^{3/2}(Y_i - \bar{y})(Y_j - \bar{y})(Y_k - \bar{y})$. On the plausible assumption that R is unimportant when n is large, we find that the skewness coefficient of the conditional distribution of h is approximately

$$\frac{2\sqrt{2} \sum_i (Y_i - \bar{y})^3}{[\sum_i (Y_i - \bar{y})^2]^{3/2}}, \tag{61}$$

and this will be small if the fitted values have a large spread and/or a third moment close to 0. This qualitative finding is no doubt true fairly generally.

If the evidence points to a value of χ somewhat different from 0, we may consider transforming the observations to reduce the heteroscedasticity. If all the observations are positive, a simple power transformation, say the pth power, could be used. We find that the variance of y_i^p is roughly $p^2\mu_i^{2(p-1)}$ Var (y_i), and this is roughly constant if $p = 1 - \chi\bar{\mu}/2$. So an estimate of the power required to make the error variance constant is

$$p = 1 - \frac{1}{2}h\bar{Y}. \tag{62}$$

The zeroth power, $p = 0$, is to be interpreted as the logarithmic transformation. (See Tukey [20] for a general discussion of such transformations.)

3.2. *Nonadditivity.* Tukey [18], [16], [19] has proposed a test which he has called "one degree of freedom for nonadditivity," designed to detect the following sort of departure from the ideal statistical conditions: the observed variable y

is a function of another variable x, such that the ideal conditions apply to (x_i). If we can determine what function y is of x, then by taking the inverse function of the observations (y_i) we shall obtain transformed readings satisfying the ideal conditions.

What functions shall we consider? If y were just a (nonzero constant) multiple of x, then in the absence of prior knowledge concerning the values of the parameters (θ_r) and the error variance σ^2 it would be impossible to say that the goodness of fit of the (x_i) to the hypothetical ideal conditions was any different from that of the untransformed (y_i). The same would be true if y differed from x only by an added constant, provided design condition 1 was satisfied; and if it were not, the effect of changing the origin of the y-scale could be investigated by introducing a general mean among the parameters (θ_r), after which condition 1 would be satisfied.

Supposing then condition 1 to be satisfied, we see that if y is a linear function of x, the (y_i) satisfy the ideal conditions as well as the (x_i). Only a nonlinear function is of any interest, and so let us suppose that $y = x + \varphi(x - \mu_0)^2$, where φ is a constant close to zero and μ_0 is some convenient central value. We are going to assume that the (x_i) are realizations of independent chance variables, such that x_i has the normal distribution $N(\mu_i, \sigma^2)$, where $\mu_i = \sum_r a_{ir}\theta_r$, the matrix **A** being given.

Let **Q** be defined as before in terms of **A**, and let (z_i) denote, as before, the residuals when the least-squares method is applied to (y_i), as though φ were zero. We have

$$z_i = \sum_j q_{ij}y_j = \sum_j q_{ij}x_j + \varphi \sum_j q_{ij}(x_j - \mu_0)^2. \tag{63}$$

Hence, remembering condition 1, we have

$$E(z_i) = \varphi \sum_j q_{ij}[(\mu_j - \mu_0)^2 + \sigma^2] = \varphi \sum_j q_{ij}\mu_j^2. \tag{64}$$

Provided **Q** is not such that the right side vanishes identically, we can say that the (z_i) have a linear regression on $(\sum_j q_{ij}\mu_j^2)$, and we shall therefore study the statistic

$$\sum_{ij} z_i q_{ij} Y_j^2 = \sum_i z_i Y_i^2. \tag{65}$$

Let

$$\begin{aligned} A_i &= \sum_j q_{ij}(x_j - \mu_0), & B_i &= \sum_j q_{ij}(x_j - \mu_0)^2, \\ C_i &= \sum_j p_{ij}(x_j - \mu_0), & D_i &= \sum_j p_{ij}(x_j - \mu_0)^2, \end{aligned} \tag{66}$$

where $p_{ij} = \delta_{ij} - q_{ij}$. Then

$$\begin{aligned} \sum_i z_i Y_i^2 = \sum_i z_i(Y_i - \mu_0)^2 &= \sum_i (A_i + \varphi B_i)(C_i + \varphi D_i)^2 \\ &= \sum_i [A_i C_i^2 + \varphi(2A_i C_i D_i + B_i C_i^2)] + O(\varphi^2). \end{aligned} \tag{67}$$

The moment-generating function of the joint distribution of A_i, B_i, C_i, D_i can be written down without difficulty, in view of the independent normality of the $(x_j - \mu_0)$. We find $E(A_iC_i^2) = 0$,

$$E(A_iC_iD_i) = 2\sigma^4 \sum_j q_{ij}(p_{ij})^2 + 2\sigma^2\mu_i \sum q_{ij}p_{ij}\mu_j, \tag{68}$$

and an expression for $E(B_iC_i^2)$ which leads to

$$E(\sum_i z_iY_i^2) = \varphi[\sum_{ij} q_{ij}\mu_i^2\mu_j^2 + \sigma^2 \sum_{ij} q_{ij}p_{ii}\mu_j^2 + 8\sigma^2 \sum_{ij} q_{ij}p_{ij}\mu_i\mu_j + 6\sigma^4 \sum_{ij} q_{ij}(p_{ij})^2] + O(\varphi^2). \tag{69}$$

Provided always that $\mathbf{Q}$ does not annihilate the vector (μ_i^2), the first term inside the brackets is more important than the others when the (μ_i) are substantially different from each other. (The second term vanishes under condition 2.) The following rough estimate of φ is therefore suggested:

$$f = \frac{\sum_i z_iY_i^2}{\sum_{ij} q_{ij}Y_i^2Y_j^2}. \tag{70}$$

It would be possible to define an estimate f^* with modified denominator, analogous to the h^* defined by (52), but we refrain.

Given an estimate f of φ, if all the observations are positive, we might consider transforming the (y_i) by taking their pth powers, in order to approach closer to the ideal conditions. The power required would be estimated roughly at

$$p = 1 - 2f\overline{Y}. \tag{71}$$

Formulas (62) and (71) should be regarded as no more than an aid to approximation. If we are lucky, both will point in the same direction.

In order to make a significance test of the deviation of f from 0, it is only necessary to note that, when $\varphi = 0$, (z_i) is independent of (Y_i) and so of $(\sum_j q_{ij}Y_j^2)$. The latter is a vector in $\overline{A}$, and, provided it is not null, we may project (z_i) onto it. We see that

$$\frac{(\sum_i z_iY_i^2)^2}{\sum_{ij} q_{ij}Y_i^2Y_j^2} \tag{72}$$

can be taken out of the residual sum of squares, $\sum_i z_i^2$, as a one-degree-of-freedom component, leaving an independent remainder of $\nu - 1$ degrees of freedom, with which it can be compared. This is Tukey's exact test for nonadditivity.

Computation of the denominator of (70) or (72) does not offer special difficulty, since it is just the residual sum of squares that is obtained by the usual analysis of variance procedure when the observations (y_i) are replaced by (Y_i^2). For the row-column cross-classification we find

$$\sum_i z_iY_i^2 = \frac{2}{kl} \sum_i z_i \text{ (row total)(column total)}, \tag{73}$$

meaning that each residual is multiplied by the total of the row and the total of the column in which it occurs in the cross-classification, and

$$\sum_{ij} q_{ij} Y_i^2 Y_j^2 = \frac{4}{kl} \text{SS(rows) SS(columns)}. \tag{74}$$

4. Justification of the criteria

4.1. The sample moments or k-statistics of a simple homogeneous sample are uniquely determined by the conditions that they are symmetric functions of the observations (demonstrably a desirable property) and are unbiased estimates for arbitrary parent distributions. No such simple argument seems to be available for the statistics proposed above.

Consider the estimation of the skewness coefficient γ_1, or of the third moment $\gamma_1\sigma^3$, of the error distribution (assumed common). Why should we take the unweighted sum of cubes, $\sum_i z_i^3$, as our basic statistic, rather than some other homogeneous cubic polynomial in (z_i)? If it happens that condition 2 is satisfied and in addition every row of $\mathbf{Q}$ is a permutation of every other row, the residuals are all equally informative concerning the shape of the error distribution, and a considerable amount of symmetry may be expected to appear in an optimum test criterion. Otherwise, it is not obvious that much symmetry can be expected.

Consider the following example, in which conditions 1 and 2 are not satisfied: $n = 3$, $\nu = 2$, $E(y_i) = (i - 1)\theta$, and the errors are independently and identically distributed with zero means. We have

$$\mathbf{Q} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0.8 & -0.4 \\ 0 & -0.4 & 0.2 \end{pmatrix}. \tag{75}$$

We find that z_1 has mean 0, variance σ^2, skewness coefficient γ_1; while z_3 is independent of z_1 and has mean 0, variance $\sigma^2/5$ and skewness coefficient $-(7/\sqrt{125})\gamma_1$; and $z_2 = -2z_3$ identically. If we restrict attention to unbiased estimates of $\gamma_1\sigma^3$ of the form $\sum_i w_i z_i^3$, and choose the constants (w_i) so as to minimize the variance for the case of a normal error distribution, we obtain the estimate

$$\frac{125}{174}(z_1^3 - 7z_3^3) \qquad \text{or equivalently} \qquad \frac{125}{174}\sum_i z_i^3. \tag{76}$$

But if to this estimate the quantity

$$\frac{25}{174} z_1 z_3 (7z_1 - 25z_3) \tag{77}$$

is added, the mean is unchanged, because z_1 and z_3 are independent and have zero means, but the variance under normality is multiplied by the factor 4/5. Thus $\sum_i z_i^3$ is not the best cubic polynomial in (z_i) to use. (That adding terms in $z_1^2 z_3$ and $z_1 z_3^2$ can reduce the variance I find astonishing.)

If we modify the example by adding a fourth observation, so that $n = 4$, $\nu = 3$, and we still have $E(y_i) = (i - 1)\theta$, we obtain the following results. The unbiased estimate of the third moment of the error distribution, based on the unweighted sum $\sum_i z_i^3$, is (approximately) $0.493\sum_i z_i^3$, of which the variance if the error distribution is normal is $6.59\sigma^6$. The variance under normality of an unbiased estimate of the form $\sum_i w_i z_i^3$ is minimized for the estimate

$$0.365z_1^3 + 0.573z_2^3 + 1.015z_3^3 + 2.404z_4^3, \tag{78}$$

of which the variance if the error distribution is normal is $5.48\sigma^6$. I have not determined the cubic polynomial estimate with minimum variance.

Consider now data having a one-way classification, that is, consisting of several homogeneous samples of possibly unequal size, from each of which we estimate a mean. Condition 1 is satisfied but not condition 2, in general. Let the rth sample be of size n_r, so that $\sum_r n_r = n$ and $\sum_r (n_r - 1) = \nu$. Let $\sum_{i(r)}$ denote a summation over the values of i for the rth sample. From each sample separately we can estimate $\gamma_1\sigma^3$ by the sample moment

$$\frac{n_r^2 \sum_{i(r)} z_i^3}{(n_r - 1)(n_r - 2)}, \tag{79}$$

of which the variance under normality is $6n_r^2\sigma^6/(n_r - 1)(n_r - 2)$. The linear combination of these estimates that has minimum variance under normality is immediately seen to be a constant multiplied by the unweighted sum $\sum_i z_i^3$, which is thus the best estimate of the form $\sum_i w_i z_i^3$, and is indeed the best cubic polynomial estimate (as can be shown without difficulty).

More generally, it is easy to show that the unweighted sum $\sum_i z_i^3$ is the best statistic (in the above sense of minimum variance) from the class of weighted sums $\sum_i w_i z_i^3$, provided that the vector (q_{ii}) lies in A, which it does when conditions 1 and 2 hold.

The statistic $\Sigma_i z_i^3$ can be derived by a likelihood function argument, as follows. Express the common error distribution by a Gram-Charlier-Edgeworth expansion, differentiate the logarithm of the likelihood function with respect to γ_1 and then set γ_1 and all other shape coefficients zero, and replace the parameters (θ_r) and σ^2 by their maximum likelihood estimates. The resulting expression contains $\sum_i z_i^3$, as well as the lower moments $\sum_i z_i$ and $\sum_i z_i^2$. Thus $\sum_i z_i^3$ is suggested as a suitable criterion for testing whether $\gamma_1 = 0$. But the suggestion does not carry much weight unless the likelihood function is closely proportional to a normal density function, which may be expected to be the case only when there is a large amount of information about every parameter θ_r, so that ν/n is close to 1 and q_{ij} is close to δ_{ij} for all i and j. The residuals then have nearly equal variance and are nearly uncorrelated.

4.2. It is not true that all possible information about the shape of the error distribution is contained in the residuals. This is illustrated by the case of a row-column cross-classification with $k(>2)$ rows and only 2 columns. The distribution of each residual is symmetrical, the sum of cubes of residuals vanishes

identically, and it would seem that no estimate of γ_1 can be obtained from the residuals. Let the observations be denoted by y_{uv}, where $u = 1, 2, \cdots, k$, and $v = 1, 2$. Then the following expression

(80)
$$\frac{2k}{3(k-1)(k-2)} \sum_u \left[(y_{u1} - \bar{y}_1)^3 + (y_{u2} - \bar{y}_2)^3 - \frac{1}{4}(y_{u1} + y_{u2} - \bar{y}_1 - \bar{y}_2)^3 \right],$$

where $\bar{y}_v = k^{-1}\sum_u y_{uv}$, is an unbiased estimate of $\gamma_1\sigma^3$ whose variance depends on the differences between row means. It is analogous to the variance estimates given by Grubbs [15] and Ehrenberg [8], for the assumption that the errors are normally distributed with unequal variances in the two columns.

4.3. By analogy with a simple homogeneous sample, one might expect that an easier test of skewness than the g_1 statistic could be obtained by comparing the number of positive and negative signs among the residuals. But the following argument suggests that such a test would be ineffectual. For typical factorial designs, if n is large and ν somewhat less than n, it seems that the numbers of positive and negative coefficients in any column of $\mathbf{Q}$ are roughly equal, whereas for a homogeneous sample they are extremely unequal. A particular and important case of skewness occurs when one error is very much larger in magnitude than all the others. For the factorial experiment this state of affairs will not be revealed by a sharp inequality in the number of positive and negative signs of the residuals.

4.4. The above discussion has concentrated on skewness and the g_1 statistic, because that topic is conceptually easiest. For the one-way classification, as defined above, we can establish the optimality of a fictitious statistic close to h, as an estimate of the heteroscedasticity parameter χ. From each sample the variance is best estimated by $\sum_{i(r)} z_i^2/(n_r - 1)$, and if we assume that this has a linear regression on $(\mu_r - \bar{\mu})\sigma^2$, fictitiously supposed known, it is easy to show that the minimum variance estimate of χ is

(81)
$$\frac{\sum_r (\mu_r - \bar{\mu}) \sum_{i(r)} z_i^2}{\sum_r (n_r - 1)(\mu_r - \bar{\mu})^2\sigma^2},$$

and we obtain h when we replace the quantities that are in fact unknown in this expression by their obvious estimates. For the one-way classification, the g_2 statistic is more difficult to investigate than the g_1 statistic, and the f statistic for nonadditivity is not defined.

5. Examples

5.1. *A typical factorial experiment.* Yates [21] illustrated the procedure of analysis of variance by analyzing some observations from a factorial experiment on sugar beet. All combinations of three sowing dates (D), three spacings of rows (S) and three levels of application of sulphate of ammonia (N) were tested

in two replications. The plots were arranged in six randomized blocks, a part of the three-factor interaction being partially confounded. The design and yields of sugar (in cwt. per acre) are shown in table I, except that the arrangement within blocks has been derandomized. (Yates gives the original plan.) Blocks 1 to 3 form one replication, blocks 4 to 6 the other; $n = 54$.

If we are to examine the residuals from such an experiment to check on the appropriateness of the least-squares analysis, we must decide how many effects to estimate. From Yates' analysis, it appears that the largest effect present is the differences between blocks. All three factors, D, S, N, have "significant" main effects, but their interactions appear to be small. The gross mean square of all the observations about their mean is 56.19. If just the six block means are estimated, we have $\nu = 48$ and the residual mean square = 21.43. When the main effects of the factors are estimated in addition to the block means, we have $\nu = 42$ and the residual mean square = 16.08. If all the two-factor interactions are also estimated, we have $\nu = 30$ and the residual square = 14.46. When, finally, the three-factor interaction is also estimated, we obtain $\nu = 22$ and the residual mean square = 13.42, this being the mean square used by Yates for gauging the estimated treatment effects.

Suppose we stop short at estimating the block means and the treatment main effects. Then any nonzero interaction effects that occur will contaminate the residuals. One may surmise that such contamination, if small, will have little effect on outliers and less still on the g_1, g_2 and h statistics, and because of the sharp drop in informativeness of the residuals when ν is decreased it would be wise to err on the side of estimating few rather than many treatment effects. A very "objective" method of deciding what effects to estimate for the purpose of calculating residuals would be to decide the matter on the basis of prior expectation, before any analysis of the data. Another method would be to perform a full conventional analysis of variance, as Yates does, and then select only those effects whose estimates were substantially larger than the residual mean square, for example, those that were "significant" at the 5 per cent level. Whether this is a good rule I do not know, but it appears not unreasonable. Perhaps for the simpler factorial designs, not highly fractionated or "saturated," a good compromise procedure would be to estimate the block means and treatment main effects in any case, and in addition any substantial or "significant" interactions revealed by the usual analysis of variance.

In table I are shown fitted values and residuals for two analyses: (a) only block means estimated, (b) block means and main effects estimated. For (b) these are also shown graphically in figure 1. Analysis (b) is what is indicated by the above compromise procedure. Consideration is also given below to a third analysis: (c) block means, main effects and two-factor interactions estimated. But it has not seemed worthwhile to calculate the residuals for (c). For analysis (a), the residuals are contaminated by the main effects of the treatments, which seem to be appreciable, and it is noteworthy that the largest residual, -12.1 in block 1, shrinks to the much less conspicuous value -5.9

TABLE I

YIELDS AND RESIDUALS OF A FACTORIAL EXPERIMENT

Block	Treatments			Yield	Analysis (a)		Analysis (b)	
	D	S	N	(y_i)	Fitted Value $(Y_i - \bar{y})$	Residual (z_i)	Fitted Value $(Y_i - \bar{y})$	Residual (z_i)
1	0	0	2	50.5	7.3	3.2	10.8	−0.3
	1	0	0	47.8		0.5	6.5	1.2
	2	0	1	46.0		−1.3	7.5	−1.5
	0	1	0	44.6		−2.7	7.1	−2.6
	1	1	1	52.7		5.4	9.9	2.7
	2	1	2	52.2		4.9	7.7	4.4
	0	2	1	51.4		4.1	7.5	3.8
	1	2	2	45.4		−1.9	7.2	−1.8
	2	2	0	35.2		−12.1	1.1	−5.9
2	0	0	0	47.8	8.8	−1.1	8.7	−1.0
	1	0	1	52.5		3.6	11.5	1.0
	2	0	2	44.1		−4.8	9.3	−5.2
	0	1	1	49.7		0.8	12.1	−2.4
	1	1	2	49.3		0.4	11.7	−2.5
	2	1	0	47.1		−1.8	5.6	1.4
	0	2	2	56.0		7.1	9.4	6.6
	1	2	0	47.2		−1.7	5.1	2.1
	2	2	1	46.2		−2.7	6.1	0.1
3	0	0	1	45.7	−0.2	5.8	3.1	2.5
	1	0	2	43.0		3.1	2.8	0.2
	2	0	0	38.0		−1.9	−3.3	1.3
	0	1	2	50.9		11.0	3.4	7.5
	1	1	0	37.1		−2.8	−0.9	−2.0
	2	1	1	38.2		−1.7	0.1	−1.9
	0	2	0	35.4		−4.5	−3.3	−1.4
	1	2	1	36.5		−3.4	−0.5	−3.1
	2	2	2	34.2		−5.7	−2.7	−3.2
4	0	0	1	39.4	−6.1	5.5	−2.8	2.2
	1	0	0	36.4		2.5	−6.8	3.2
	2	0	2	29.9		−4.0	−5.6	−4.5
	0	1	0	33.3		−0.6	−6.2	−0.5
	1	1	2	33.6		−0.3	−3.2	−3.3
	2	1	1	35.3		1.4	−5.9	1.1
	0	2	2	31.9		−2.0	−5.6	−2.6
	1	2	1	34.4		0.5	−6.4	0.8
	2	2	0	31.3		−2.6	−12.3	3.5
5	0	0	2	33.6	−5.1	−1.3	−1.6	−4.8
	1	0	1	41.8		6.9	−2.5	4.2
	2	0	0	33.2		−1.7	−8.3	1.5
	0	1	1	36.6		1.7	−1.8	−1.6
	1	1	0	33.0		−1.9	−5.9	−1.2
	2	1	2	41.4		6.5	−4.7	6.0
	0	2	0	25.7		−9.2	−8.2	−6.1
	1	2	2	31.4		−3.5	−5.2	−3.4
	2	2	1	37.6		2.7	−7.9	5.5

TABLE I (Continued)

Block	Treatments D	S	N	Yield (y_i)	Analysis (a) Fitted Value ($Y_i - \bar{y}$)	Analysis (a) Residual (z_i)	Analysis (b) Fitted Value ($Y_i - \bar{y}$)	Analysis (b) Residual (z_i)
6	0	0	0	39.4	−4.7	4.1	−4.9	4.2
	1	0	2	43.1		7.8	−1.8	4.9
	2	0	1	26.6		−8.7	−4.5	−8.9
	0	1	2	36.0		0.7	−1.2	−2.8
	1	1	1	34.2		−1.1	−2.1	−3.8
	2	1	0	33.5		−1.8	−7.9	1.4
	0	2	1	34.9		−0.4	−4.5	−0.7
	1	2	0	32.4		−2.9	−8.5	0.8
	2	2	2	37.7		2.4	−7.3	4.9

under analysis (b). The second largest residual under analysis (a), 11.0 in block 3, becomes the second largest residual, 7.5, under analysis (b). Had there been a gross error in one of the observations, it would have been expected to appear as an outlier in both analyses.

Outlier rejection criteria can be calculated from the approximate formula (11.3) of [3], which expresses the premium charged by the rule. For a 2 per cent premium and no prior information about σ^2, we obtain for analysis (a) the follow-

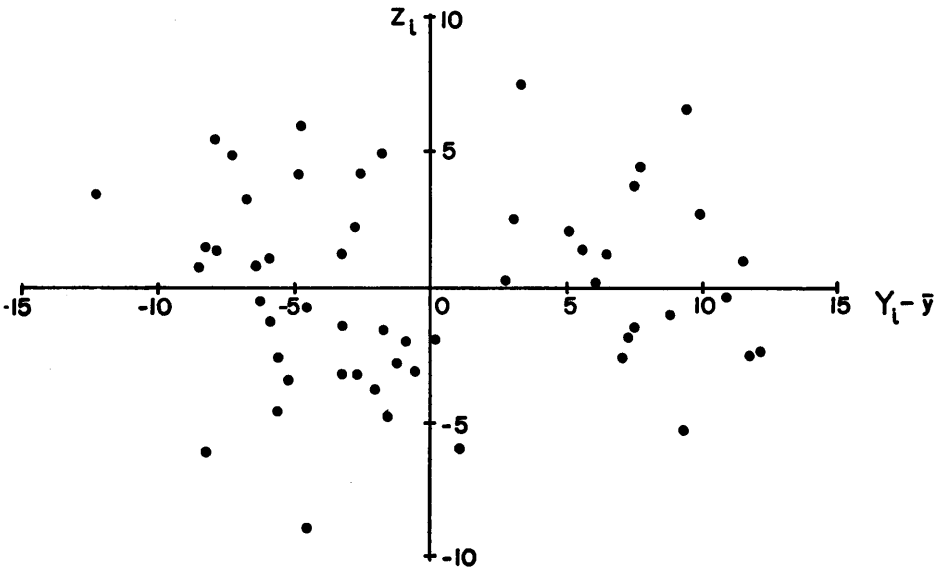

FIGURE 1

Fitted values and residuals from table I, analysis (b).

ing critical value for $\max_i |z_i|$: $2.86s = 13.2$. For analysis (b) the critical value is $2.69s = 10.8$. No rejections are indicated.

In order to calculate the formulas given in sections 2 and 3, we begin by determining **Q**. Conditions 1 and 2 are satisfied, and the rows of **Q** are permutations of each other. For analysis (a), each row of **Q** contains the following elements: 8/9 (once, in the principal diagonal), −1/9 (eight times), 0 (45 times). For analysis (b), each row of **Q** contains: 7/9 (in the diagonal), −1/9 (seven times), 1/18 (14 times), −1/18 (14 times), 0 (18 times). For analysis (c), each row of **Q** contains: 5/9 (in the diagonal), −1/3 (once), −1/9 (twice), 1/18 (18 times), −1/18 (18 times), 0 (14 times). [Check by sum and by sum of squares, using (7) and (8).] Hence we easily obtain from (18) and (37) the variances under normality of g_1 and g_2, as listed in table II. The variances for

TABLE II

VARIANCES OF STATISTICS CALCULATED FROM TABLE I

Analysis	Variance of g_1 Given by (18) and Equivalent n	Variance of g_2 Given by (37) and Equivalent n	Variance of h Given by (53)
(a)	0.142 (39)	0.600 (35)	0.00111
(b)	0.210 (26)	1.011 (19)	0.00126
(c)	0.698 (6)	3.324 (6)	0.00259

analysis (a) are in error in the sense that we have good reason to think that the residuals are contaminated by substantial treatment effects, but they are quoted to show how the variances given by the formulas change with ν. The "equivalent n" shown in brackets is that size of homogeneous sample which has most nearly the same variance of g_1 and g_2, according to (20) and (38).

In order to calculate h and Var $[h|(Y_i)]$, we first obtain a convenient expression for $\sum_{ij}(q_{ij})^2(Y_i - \bar{y})(Y_j - \bar{y})$. For analysis (a) we can use (56), obtaining

$$\sum_{ij} (q_{ij})^2(Y_i - \bar{y})(Y_j - \bar{y}) = \frac{8}{9}\,\mathrm{SS(blocks)} = 1733.67. \tag{82}$$

To find a similar expression for analysis (b), we express the vector $(Y_i - \bar{y})$ as the sum of four orthogonal vectors in A, the first consisting of the estimated block effects (differences of block means from the general mean), the other three of the estimated main effects of the factors, D, S, N, separately. We then transform each component vector by the matrix $((q_{ij})^2)$, and note that the result is orthogonal to each of the other three component vectors. Hence the required sum of squares can be obtained easily by considering the four components one at a time, as if the other three were not present, and summing the results. We get

$$\begin{aligned}\sum_{ij} (q_{ij})^2(Y_i - \bar{y})(Y_j - \bar{y}) &= \frac{2}{3}\,\mathrm{SS(blocks)} + \frac{11}{18}\,\mathrm{SS(main\ effects\ of}\ D, S, N) \\ &= 1515.62.\end{aligned} \tag{83}$$

The same method, a little less easily, yields for analysis (c)

(84)

$$\sum_{ij} (q_{ij})^2 (Y_i - \bar{y})(Y_j - \bar{y}) = \frac{1}{3}\,\mathrm{SS(blocks)} - \frac{1}{9}\,\mathrm{SS(replications)} + \frac{23}{54}\,\mathrm{SS(main\ effects\ of}\ D, S, N) + \frac{7}{18}\,\mathrm{SS(two\text{-}factor\ interactions\ of}\ D, S, N) = 724.58.$$

Hence from (53) we obtain the variances listed in table II.

For analysis (b), inspection of figure 1 does not reveal any peculiarity in the distribution of the residuals nor any suggestion of regression of (z_i^2) on (Y_i) nor of (z_i) on (Y_i^2). Calculations yield the following results, where the number following the $\pm$ sign is the standard deviation of the sampling distribution of the statistic under the ideal conditions.

(85)

$$\begin{aligned} g_1 &= -0.02 \pm 0.46, \\ g_2 &= -0.57 \pm 1.01, \\ h &= -0.023 \pm 0.035, \\ f &= 0.019 \pm 0.018. \end{aligned}$$

Tukey's test: compare the nonadditivity term 18.09 (1 degree of freedom) with the residual mean square 16.03 (41 degrees of freedom), ratio = 1.13.

In the absence of further development of theory, it is reasonable to regard the above standard deviations as crude approximations to standard errors of estimation of the hypothetical parameters γ_1, γ_2, χ, φ. They are very large, and the estimates are consequently very poor. Consider h, for example. The fitted values (Y_i) cover a range of about 24. The factor by which the error standard deviations (σ_i) change from one end of this range to the other is exp (12χ), on the regression hypothesis of section 3.1. Two standard errors above and below h give us 0.05 and -0.09, roughly, limits between which we may hope that χ lies, and the corresponding limits for exp (12χ) are 1.8 and 1/3.1. Thus although there is no evidence of heteroscedasticity, the data are reasonably compatible with a change in the error standard deviation by a factor of 2 in either direction over the range covered by (Y_i). A much larger body of data would be needed for usefully precise estimation.

One point in the above discussion merits further consideration. The variances in table II have been presented merely to illustrate the formulas. When we have decided which analysis to use, we can proceed as indicated above. Significance tests of the deviations of g_1, g_2, h, from 0 are in any case valid tests of the ideal statistical conditions. But if we wish to compare the sensitivities of the three analyses, (a), (b), (c), for detecting nonnormality or heteroscedasticity, further

thought is needed, because what each statistic, g_1, g_2, h, estimates differs for each analysis, in accordance with the difference in the implied ideal conditions.

Consider the h statistic. Let us suppose, for the purpose of comparison, that in fact the (μ_i) are linear combinations of block means, main effects and two-factor interactions of the factors, but there is no three-factor interaction nor any other effects on the means; and suppose further that there is a small regression of error variance on the mean, with parameter χ. Then for analysis (c), we obtain (to the best of our knowledge) a nearly unbiased estimate of χ by calculating h_c^*, which turns out to be $1.202\, h_c$. (We use the suffices a, b, c here to distinguish between analyses.) But in analyses (a) and (b), some real treatment effects are left in the apparent error variation, and the result of this seems to be roughly, on the average, to increase the apparent residual variance by a constant amount, independent of the mean, and therefore to diminish the apparent magnitude of χ. Hence the following roughly unbiased estimates of χ are suggested:

$$\frac{s_a^2}{s_c^2} h_a^* = 1.565\, h_a, \qquad \frac{s_b^2}{s_c^2} h_b^* = 1.197\, h_b. \tag{86}$$

The variance of each estimate is presumably roughly found by multiplying the

TABLE III

AN EXPERIMENT WITH INSECTICIDES

The three entries in each cell are the observed count of leatherjackets y_{uv}, the fitted value Y_{uv} in parentheses, and the residual z_{uv}

	Treatments					
Block	1(control)	2(control)	3	4	5	6
1	92 (77.2) 14.8	66 (77.2) −11.2	19 (35.9) −16.9	29 (23.8) 5.2	16 (18.9) −2.9	25 (13.9) 11.1
2	60 (63.9) −3.9	46 (63.9) −17.9	35 (22.6) 12.4	10 (10.4) −0.4	11 (5.6) 5.4	5 (0.6) 4.4
3	46 (67.9) −21.9	81 (67.9) 13.1	17 (26.6) −9.6	22 (14.4) 7.6	16 (9.6) 6.4	9 (4.6) 4.4
4	120 (77.2) 42.8	59 (77.2) −18.2	43 (35.9) 7.1	13 (23.8) −10.8	10 (18.9) −8.9	2 (13.9) −11.9
5	49 (65.5) −16.5	64 (65.5) −1.5	25 (24.3) 0.7	24 (12.1) 11.9	8 (7.3) 0.7	7 (2.3) 4.7
6	134 (86.9) 47.1	60 (86.9) −26.9	52 (45.6) 6.4	20 (33.4) −13.4	28 (28.6) −0.6	11 (23.6) −12.6

corresponding entry in table II by the square of the multiplier of h. We obtain

(87) (a) 0.0027, (b) 0.0018, (c) 0.0037.

While these estimated variances are crude, they are no doubt correct in indicating that analysis (b) is the most sensitive, and analysis (c) the least, for detecting a departure of χ from 0.

5.2. *Insect counts.* To demonstrate that with even a small body of data the methods of this paper are capable of revealing a gross enough violation of the ideal conditions, let us consider some leatherjacket counts which Bartlett [5] quoted as an example to illustrate the use of a transformation of the data in reducing heteroscedasticity. In each of six randomized blocks (replications) there were six plots, four treated by various toxic emulsions and two untreated as controls. In table III are shown the total counts of leatherjackets recovered on each plot, together with the fitted values and residuals when treatment means and block means are estimated. Let the observation in the uth row and vth column of table III be denoted by y_{uv}, with $u, v = 1, 2, \cdots, 6$. Then the fitted values are given by

$$Y_{u1} = Y_{u2} = \frac{1}{6}\sum_v y_{uv} + \frac{1}{12}\sum_u (y_{u1} + y_{u2}) - \bar{y},$$

(88)

$$Y_{uv} = \frac{1}{6}\sum_v y_{uv} + \frac{1}{6}\sum_u y_{uv} - \bar{y}, \qquad v \geqq 3.$$

The analysis of variance goes as shown in table IV.

TABLE IV

ANALYSIS OF VARIANCE FOR EXPERIMENT WITH INSECTICIDES

	Degrees of Freedom	Sums of Squares	Mean Squares
Blocks	5	2358.22	471.64
Treatments	4	24963.14	6240.78
Residual	26	8502.53	327.02

Because there are twice as many control plots as of each type of treated plot, condition 2 is violated. Twelve rows of **Q** contain the elements: 7/9 (in the diagonal), −2/9 (once), −5/36 (four times), −1/18 (10 times), 1/36 (20 times). The other 24 rows of **Q** contain: 25/36 (in the diagonal), −5/36 (10 times), 1/36 (25 times). It is straightforward to calculate the various functions on **Q** that are needed.

$$\sum_i (q_{ii})^2 = 18.833, \qquad \sum_{ij} (q_{ij})^2 = 12.778,$$

$$D = 8.649, \qquad F = 0.034,$$

(89)

$$\sum_{ij} (q_{ij})^2 (Y_i - \bar{Y})(Y_j - \bar{Y}) = \frac{11}{18}\,\mathrm{SS(blocks)} + \frac{5}{9}\,\mathrm{SS(treatments)} + \frac{25}{26}\left[\frac{1}{12}\sum_u (y_{u1} + y_{u2}) - \bar{y}\right]^2.$$

Hence we find

$$\begin{aligned} g_1 &= 1.77 \pm 0.61, \\ g_2 &= 4.39 \pm 1.35, \\ h &= 0.046 \pm 0.011, \\ f &= 0.0135 \pm 0.0071. \end{aligned} \tag{90}$$

Tukey's test: compare the nonadditivity term 1196.54 (one degree of freedom) with the residual mean square 292.24 (25 degrees of freedom), ratio = 4.09, about the upper 5.1 per cent point of the variance-ratio distribution.

The power transformation suggested by substituting the above value of h into (62) is $p = 0.12$, and by substituting f into (71) is $p = -0.04$. Thus h

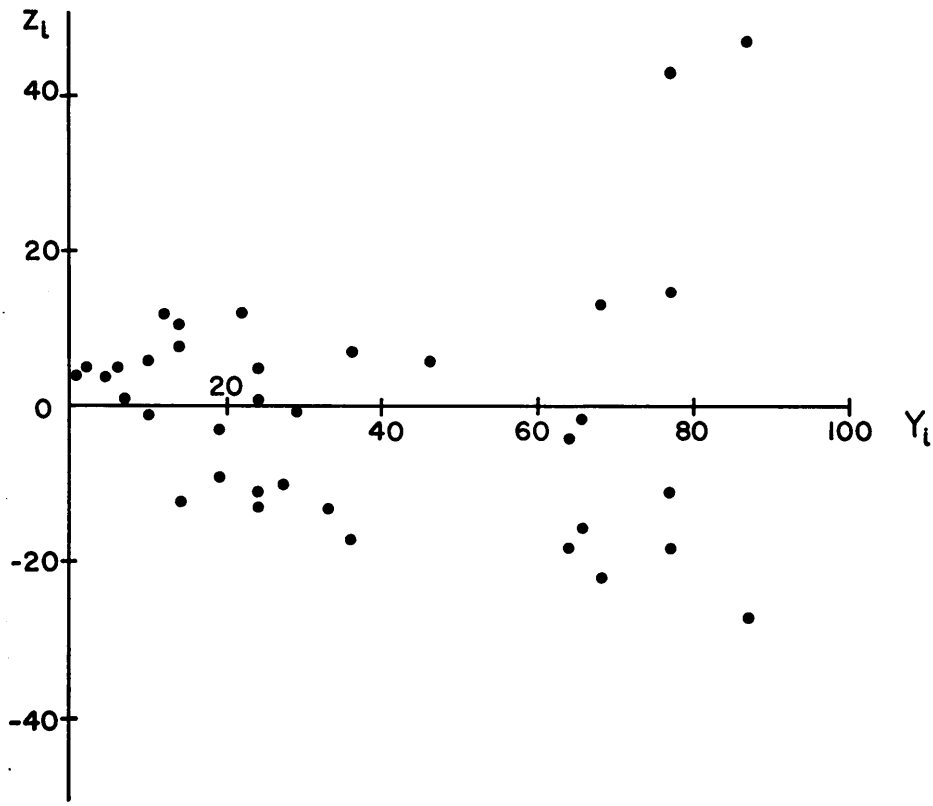

FIGURE 2

Fitted values and residuals from table III.

and f both indicate something like a logarithmic transformation of the original counts. This is satisfactory, because experience with insect counts suggests that under homogeneous conditions they will closely approximate a negative binomial distribution (see Evans [9]), and that the negative binomial distributions for counts made with similar technique under different treatments or conditions may be expected to have roughly the same exponent. If k denotes the presumed common exponent, the transformed variables

$$\log\left(y + \frac{1}{2}k\right) \quad \text{or} \quad 2\sinh^{-1}\left(\frac{y + \frac{3}{8}}{k - \frac{3}{4}}\right)^{1/2} \tag{91}$$

will have almost constant variance equal to $\psi'(k)$, the second derivative of the logarithm of the Gamma function at k, and roughly normal distribution, provided that $E(y)$ and k are not too small. We can estimate k by comparing the mean squares of the counts for each treatment with the mean count for that treatment, ignoring block differences; see Bliss and Owen [6], who also give an

TABLE V

TRANSFORMED COUNTS $\log(y + 4)$ FROM TABLE III

Block	1(control)	2(control)	3	4	5	6
1	4.56	4.25	3.14	3.50	3.00	3.37
2	4.16	3.91	3.66	2.64	2.71	2.20
3	3.91	4.44	3.04	3.26	3.00	2.56
4	4.82	4.14	3.85	2.83	2.64	1.79
5	3.97	4.22	3.37	3.33	2.48	2.40
6	4.93	4.16	4.03	3.18	3.47	2.71
Mean transformed counts	4.29		3.52	3.12	2.88	2.50

analysis of these observations. We obtain a pooled estimate of about 5 or 6 for k. This estimate may be expected to be on the low side, because block differences have been ignored. Let us therefore guess the round figure of 8 for k, and consider the transformed counts $\log(y + 4)$, shown in table V. The analysis of variance now goes as shown in table VI. The residual mean square is equal

TABLE VI

ANALYSIS OF VARIANCE FOR TRANSFORMED COUNTS

	Degrees of Freedom	Sums of Squares	Mean Square
Blocks	5	1.3145	0.2629
Treatments	4	16.3918	4.0980
Residual	26	3.2692	0.1257

to $\psi'(8.44)$ approximately, or about 94 per cent of $\psi'(8)$, so the guessed value for k has been quite well confirmed. The residuals after fitting block and treatment means to the entries in table V do not show any interesting phenomena, which is what one would expect with so few observations; and they are not reproduced here. One may conclude that the scale of $\log(y + 4)$ is satisfactory for viewing the counts through the simple row-column least-squares analysis. (Bliss and Owen recommend for these counts the transformation $\log(y + 12.6)$, for reasons that do not entirely convince. Of course, almost any logarithmic transformation will lead to apparently well-behaved residuals, with so few observations.)

5.3. *Comparisons of designs.* The informativeness of the residuals depends on $\mathbf{Q}$, which in turn depends largely but not entirely on the values of n and ν. It is possible for two designs to have the same n and ν and yet differ perceptibly in the properties of their residuals. This will be illustrated by two examples with very small n.

Consider first the estimation of a quadratic response surface representing the dependence of mean yield on the (continuously variable) levels of two factors. The rotatable designs of Box and Hunter [7] do not in general satisfy condition 2, but it can happen that they do, and there is something to be said for trying to secure this if possible. The designs are specified by points in the factor-level space, here a plane, representing treatment combinations at which observations are made. One suitable design consists of two independent observations at a center point and ten further observations, one at each of ten points spaced equally on the circumference of a circle round the center point. With all the observations arranged in one randomized block, we have $n = 12$, $\nu = 6$, and

(92)

$$\mathbf{Q} = \begin{pmatrix} 0.5 & -0.5 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -0.5 & 0.5 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.5 & -0.324 & 0 & 0.124 & 0 & -0.1 & 0 & 0.124 & 0 & -0.324 \\ 0 & 0 & -0.324 & 0.5 & -0.324 & 0 & 0.124 & 0 & -0.1 & 0 & 0.124 & 0 \\ 0 & 0 & 0 & -0.324 & 0.5 & -0.324 & 0 & 0.124 & 0 & -0.1 & 0 & 0.124 \\ & & & & \cdot & \cdot & \cdot & \cdot & & & & \end{pmatrix}$$

Rows 3 to 12 of $\mathbf{Q}$ are permutations of each other; the number -0.324 stands for $-(\sqrt{5} + 1)/10$ and 0.124 for $(\sqrt{5} - 1)/10$. For this design we find

$$\text{Var}(g_1) = 4.5, \qquad \text{Var}(g_2) = 10.05. \tag{93}$$

Another possible design, also satisfying condition 2, consists of two observations at the center point and two at each of five points spaced equally on the circumference of the circle. Now every row of $\mathbf{Q}$ contains the elements 0.5 (in the diagonal), -0.5 (once), 0 (10 times), like the first two rows above, and we find that g_1 is not defined but g_2 is a better estimate than before,

$$\text{Var}(g_2) = 5.76. \tag{94}$$

For the purpose of detecting outliers, the first of these designs is better than the second (neither is good).

As another example, consider the estimation of the main effects of eight two-level factors, A, B, C, D, E, F, G, H, by a fractional factorial design arranged in one block, so that $n = 16$ and $\nu = 7$. One possibility is to use the alias subgroup generated by the interactions $ABCD$, $CDEF$, $ACEG$, $EFGH$; another possibility is to use the alias subgroup generated by ABC, CDE, EFG, AGH. For the first design, every row of $\mathbf{Q}$ contains the elements: $7/16$ (in the diagonal and also once elsewhere), $-1/16$ (14 times). For the second design, each row of $\mathbf{Q}$ has: $7/16$ (in the diagonal), $3/16$ (twice), $-3/16$ (four times), $1/16$ (four times), $-1/16$ (five times). We can make the following comparisons.

(a) For checking on *outliers*, the second design is much better, because no pair of residuals has a correlation coefficient exceeding $3/7$ in magnitude, whereas for the first design the residuals are equal in pairs.

(b) For checking *nonnormality*, the first design is better (though both are very bad), because we find

for the first design: $\text{Var}\,(g_1) = 1.20$, $\text{Var}\,(g_2) = 8.77$;

for the second design: $\text{Var}\,(g_1) = 2.64$, $\text{Var}\,(g_2) = 24.56$.

(c) Checking *heteroscedasticity* is possible only with the second design, because we find

for the first design: $\sum_j (q_{ij})^2(Y_j - \bar{y}) = 0$ for all i;

for the second design: $\sum_{ij} (q_{ij})^2(Y_i - \bar{y})(Y_j - \bar{y}) = \frac{1}{4}\text{SS}(A, C, E, G) + \frac{1}{8}\text{SS}(B, D, F, H)$.

(d) For checking *nonadditivity*, the first design is better, because no two-factor interactions are confounded with main effects, whereas for the second design 12 of the 28 possible two-factor interactions are so confounded.

Most statisticians most of the time would prefer the first design to the second, as it leaves the main effects clear of two-factor interaction aliases. If confidence were felt that all interactions were negligible, the better control on outliers might make the second design preferable.

5.4. *Designs with equal-magnitude correlations.* Some interest attaches to the possibility of designs satisfying conditions 1 and 2 and such that all the off-diagonal elements of $\mathbf{Q}$ are equal in magnitude. Such designs are especially favorable for the detection of gross errors as outliers, and permit several of the formulas of this paper to assume their simplest form.

A necessary prerequisite for such a design is that $[\nu(n-1)/(n-\nu)]^{1/2}$ be an integer, this being the difference between the numbers of negative and positive signs among the off-diagonal elements of any row of $\mathbf{Q}$. A simple homogeneous sample having $\nu = n - 1$ satisfies this condition, for any n; and the condition is also satisfied for any n if we set $\nu = 1$. Some other combinations of n and ν satisfying the condition are shown in table VII. All possibilities with $n \leqq 36$ and $1 < \nu < n - 1$ are listed, together with a few others having $n > 36$.

TABLE VII

POSSIBILITIES FOR DESIGNS WITH EQUAL-MAGNITUDE CORRELATIONS

n	ν	n	ν
9	3, 6	33	11, 22, 25, 27
10	5, 8	35	*18*
15	*8*	36	*15, 21*
16	*6*, 10**	—	
21	*16*	49	21, 28
25	10, 15	64	8, *28**, *36*, 50, 56
26	13	81	36, 45
28	7, 16, 21, 25	100	*45**, *55*

It would seem that for most of the listed combinations of n and ν there is no actual design with equal-magnitude correlations. Any ordinary type of orthogonal design has the property that every element of $\mathbf{Q}$ is an integer multiple of $1/n$. If the design has equal-magnitude correlations, we see that $[\nu(n - \nu)/(n - 1)]^{1/2}$ must be an integer. Possibilities in table VII satisfying this condition are shown with the value of ν in italics. The asterisk indicates known solutions. Solutions for $n = 16$ were given in [3]. The possibility $n = 64$ and $\nu = 28$ is realizable as a hyper-Graeco-Latin square, formed by superimposing three orthogonal 8×8 Latin squares. The possibility $n = 100$ and $\nu = 45$ is similarly realizable by superimposing four orthogonal 10×10 Latin squares, of which the existence has been demonstrated by R. C. Bose. There is an unlimited sequence of such Latin square designs having equal-magnitude correlations. In particular, they exist whenever n is a power of 4. One may conjecture that the possibility $n = 64$ and $\nu = 36$, may be realizable by a 2^6 factorial experiment with a suitable selection of interactions estimated.

5.5. *A counterexample.* Not every kind of departure from the ideal statistical conditions can be seen clearly by examining the residuals (quite apart from the imprecision arising from large sampling errors). A good example of the possible unhelpfulness of looking at residuals is provided by some data (apparently slightly faked) quoted by Graybill [14], showing the yields of four varieties of wheat grown at 13 locations in the state of Oklahoma. If the residuals from row and column means are calculated, they seem to have different mean squares in the four columns, and we might be led to modify the least-squares analysis by postulating a different error variance for each variety. But if the original observations (*not* these residuals) are examined more closely, it will appear that the locations do not have a simple additive effect, but rather the varieties seem to respond with different sensitivity to different locations. Variety number 3 has nearly the same yield at all locations, whereas the other varieties show pronounced differences, on the whole similar in sign but varying in magnitude. It is primarily the additive assumption about rows and columns which is inappropriate here, and needs to be modified. A more plausible assumption would

be the following. The observation y_{uv} in the uth row and vth column, where $u = 1, 2, \cdots, 13$, and $v = 1, 2, 3, 4$, is independently drawn from a population with mean $\theta_v + \beta_u/\alpha_v$ and variance σ^2/α_v^2. The parameters (α_v) can be estimated as inversely proportional to the root mean squares of entries in each column of the original table: we get the estimates 0.21, 0.36, 1.00, 0.42. If now the entries in each column are multiplied by the corresponding $\hat{\alpha}_v$ we obtain a set of numbers in which (as near as we can judge) row and column effects are additive and the error variance is constant.

6. Discussion

6.1. Given some observations (y_i) and associated linear hypothesis, that is, given the matrix **A**, we can group together under four main headings the various questions that can be asked concerning the appropriateness of a least-squares analysis.

(i) *Are the observations trustworthy?*

If the answer is yes, we can proceed to challenge the component parts of the statement of ideal statistical conditions.

(ii) *Is it reasonable to suppose the* (y_i) *to be realizations of independent chance variables such that there exist parameter values* (θ_r) *such that* $E(y_i) = \sum_r a_{ir}\theta_r$*?*

(iii) *Is it reasonable to suppose the* (y_i) *to be realizations of independent chance variables all having the same variance?*

(iv) *Is it reasonable to suppose the* (y_i) *to be realizations of independent chance variables all normally distributed?*

One might add a fifth query concerning the supposition of independence, but that seems to be a metaphysical matter. A phenomenon which could be thought of as one of dependence between chance variables could also be thought of in terms of independent chance variables having a different mutual relation. In many applications of the method of least squares, independence is a natural assumption, either because of the physical independence of the acts of observation, or because of a randomization of the layout. One might add a further, more radical, query about interpreting the observations as any sort of chance phenomena. Why think in terms of chance variables at all? B. de Finetti and L. J. Savage have claimed that it is possible to express all kinds of uncertainty regarding phenomena in terms of subjective probabilities. To avail ourselves of a distinct physical concept of random phenomena (here referred to by the label "chance") is, they have shown, unnecessary. But the physical concept is nevertheless attractive, both because it is philosophically simpler than any logical concept of probability, and because of its familiarity in orthodox statistical thinking. Be all this as it may, we shall here think exclusively in terms of independent chance variables. Let us now consider each of the above four types of question in turn.

The first question concerns whether we should accept the observations at their face value, or discard them, partly or wholly. If the general level of the observa-

tions, that is, $\bar{y}$, or the calculated estimates of some of the parameters (θ_r), or the estimate of the error variance σ^2, are strongly discrepant with our prior expectations, we shall suspect that a blunder has been made somewhere, in carrying out the plan of observation, or in the arithmetical reduction of the original readings. If the blunder cannot be identified and rectified, the observations will perhaps be rejected altogether. The possibility that occasionally a single observation is affected by a gross error can be allowed for by examining the largest residuals. In some circumstances it will be appropriate to adopt a definite routine rejection rule for outliers.

Under the heading (ii) comes a familiar question. In the analysis of a factorial experiment, how many interactions should be individually estimated, how many should be allowed to contribute to the estimation of the error variance? Is the matrix **A** big enough, or should further columns (representing interactions) be added, or conversely, can some columns safely be deleted? Another sort of question that can arise concerns the scale or units in which the observations can best be expressed. When what is observed is the yield of a production process, we are usually interested rather strictly in estimating (or maximizing) *mean* yields, and a nonlinear transformation of the observations might well be considered to be out of place, even if it brought some apparent advantages for the statistical analysis. But in other cases less easily resolved doubts arise about the proper scale of measurement. If electrical resistance is observed, would it be better expressed by its reciprocal, conductivity? If the dimension of objects of fixed shape is observed, should a linear dimension be recorded, or its square, or cube? In some population studies we expect treatment effects to be multiplicative, and a linear hypothesis about (μ_i) becomes more plausible after the counts have been transformed logarithmically. In recording sensory perceptions or value judgments arbitrary numerical scores are sometimes used, and on the face of it these might as well be transformed in almost any manner. We may hope that by transforming the observations we can arrange that the ideal statistical conditions obtain to a satisfactory degree of closeness, for a small parameter set (θ_r). Tukey's nonadditivity test (f statistic) is valuable as an aid to reducing the number of interactions that need to be considered.

Under the heading (iii), the h statistic is designed to show up that kind of dependence of the error variance on the mean that could be removed by a power transformation of the observations. Other possible sorts of heteroscedasticity can be detected by examining the residuals, but they are not studied here.

Question (iv) regarding nonnormality can be examined with the g_1 and g_2 statistics.

6.2. The four statistics studied in this paper, g_1, g_2, h, f, and also the largest residual, $\max_i |z_i|$, studied in [3], are by no means independent. If the ideal conditions fail in some particular respect, more than one of these statistics may respond. For example, if n is not very large and if one observation is affected by a gross error, g_1 and g_2 are likely to be large, and possibly also f and h if the affected observation has an extreme mean. Any kind of heteroscedasticity may

affect both g_2 and h. Thus it may be much easier, in a particular case, to assert that the ideal conditions do not hold, than to say what does and what ought to be done.

Certainly all five statistics are not equally important or interesting. I suggest that it is always worthwhile, if computational facilities permit, to make some sort of check for outliers. Perhaps this is the only universal recommendation that should be made. If we are willing to consider transformations, then f and h become interesting. Above we began by considering g_1 and g_2, but that was only because they were conceptually a little simpler than f and h. It seems that only from a large bulk of data, such as a whole series of experiments in a particular field, can any precise information be distilled about the shape of the error distribution. For smaller amounts of data, calculating g_1 and (especially) g_2 is a waste of time. The graphical plotting of residuals against fitted values is no doubt a good routine procedure, and can be done automatically by a computer.

6.3. *Significance tests for theoretical hypotheses.* In sections 2 and 3 above special attention was paid to the sampling distribution of the statistics under the full ideal conditions, so that significance tests of departure from the ideal conditions could be made. In [3], on the other hand, it was suggested that the traditional approach to the rejection of outliers through significance tests was inappropriate, and that choosing a rejection rule was a decision problem similar to deciding how much fire insurance to take out on one's house. The difference of approach to related problems calls for explanation. What is at issue is the relevance of significance tests in this context.

On a previous occasion [2] I have pointed to two very different situations in which a "null hypothesis" is of special interest, and some sort of test of conformity of the observations seems to be called for. In the first situation, there is a certain hypothesis which there is good reason to expect may be almost exactly true. For example, the hypothesis may be deduced from a general mathematical theory which is believed to be good, and the observations have been made to test a prediction of the theory. Another example would be an experiment on extrasensory perception; most people believe that no such thing as ESP exists and that a "null hypothesis" deduced from simple laws of chance must be true, whereas the experimenter hopes to obtain observations that do not conform with this null hypothesis. Yet another example would be a set of supposed random observations from a specified chance distribution, derived from pseudo-random numbers, where we might wish to test conformity of the observations with the nominal distribution. In such situations we wish to know whether the observations are compatible with the hypothesis considered. It is irrelevant to ask whether they might also be compatible with other hypotheses. Usually we are reluctant to try to embed the null hypothesis in a broader class of admissible hypotheses, defined in terms of only one or two further parameters, such that one of these hypotheses must be true. If the evidence shows the null hypothesis to be untenable, shows, that is, that we need to think again, we may perhaps consider patching up the hypothesis by introducing an extra parameter or two,

but we look first at some observations to see what sort of modification is needed. If indeed we had a class of admissible hypotheses at the outset, with not too many nuisance parameters, the likelihood function would be a complete summary of the observations, and we could make inferences with Bayes' theorem. But in the situation envisaged there is no small enough class of admissible hypotheses, no intelligible likelihood function, Bayesian inference is not available, and it is natural to fall back on the primitive significance test, of which Karl Pearson's χ^2 test of goodness of fit is the classic example. In such a test a criterion (function of the observations) is chosen, with an eye to its behavior under some particular alternatives considered possible, and the value of the criterion calculated from the data is compared with its sampling distribution under the null hypothesis, for a specified sampling rule. (Sometimes it is a conditional distribution that is considered.) The end result is a statement that the criterion has been observed to fall at such and such a percentile of its sampling distribution. Extreme percentiles (or more generally certain special percentiles) are regarded as evidence that the observations do not conform with the null hypothesis. This type of analysis of the data is related to a null hypothesis expressed in terms of chances in the same way, as nearly as possible, as a simple count of observed favorable and unfavorable instances is related to a universal hypothesis, of the type "all A's are B's." Such an analysis is not a decision procedure, it does not imply any decisions. We do not necessarily believe a universal hypothesis is true just because no contrary instances have been observed, nor do we necessarily abandon a universal hypothesis just because a few contrary instances have been observed. Similarly, our attitude towards a statistical hypothesis is not necessarily determined by the extremeness of the observed value of the test criterion. The significance test is evidence, but not a verdict. Its function is humble, but essential. The only way that we can see whether a statistical hypothesis (that is, a hypothesis about physical phenomena, expressed in terms of chances) is adequate to represent the phenomena is through significance tests, or, more informally, by noticing whether the observations are such as we could reasonably expect if the hypothesis were true. All scientific theories ultimately rest on a simple test of conformity: universal hypotheses are confirmed by noting the incidence of favorable cases, statistical hypotheses are confirmed by significance tests. Any proposal of a class of admissible statistical hypotheses, prerequisite for the ordinary use of Bayes' theorem, depends for its justification, if it has one, ultimately on significance tests.

The above argument constitutes, I believe, a defense of Fisher's attitude to significance tests, in his later writings. In [2] I had not realized the importance of the absence of a class of admissible hypotheses, and was therefore skeptical concerning orthodox significance tests. In addition to tests of theoretical hypotheses, discussed above, for which orthodox significance tests seem to be appropriate, and to tests of simplifying hypotheses, as discussed below, it appears that there is a third type of situation to which the name test can be reasonably

applied, as follows. There is a class of admissible hypotheses, and the problem would be an ordinary one of estimation except that the prior probability is partly concentrated on a lower dimensional subspace of the parameter space. When we come to use Bayes' theorem, the calculations are of the sort termed a significance test by H. Jeffreys. In [2] I did not perceive that inference problems of this type could indeed arise in science; convincing examples have since been given by D. V. Lindley (testing for linkage) and L. J. Savage (testing for statistic acid).

6.4. *Simplifying hypotheses.* Can it be said that the ideal statistical conditions for a least-squares analysis constitute a theoretical null hypothesis of the above sort, so that to check it we resort to significance tests? Not without some apology. We can hardly claim that we have theoretical reasons for believing the ideal conditions to hold. We have seen in section 5 that with small amounts of data it is remarkably difficult for the ideal conditions to disprove themselves. No doubt many users of the least-squares method believe that the ideal conditions are very nearly satisfied in practice. If that belief is false (in some field of observation), significance tests will eventually show it, if enough observations are made, and then the user must consider whether the discrepancies matter and what he ought to do about them. It is a common scientific practice to make bold use of the simplest hypotheses until they are clearly shown to need modification. That practice is presumably the best excuse for waiting until discrepancies with the ideal conditions are clearly visible before questioning the ordinary direct use of the method of least squares.

The hypothesis that the ideal statistical conditions are satisfied is an example of what was called in [2] a simplifying hypothesis. We are disposed to act as though we believed the hypothesis to be true, not because we really do believe it true, but because we should be so pleased if it were. Once we realize this, we see that significance tests are not strictly relevant, though possibly useful in shaking us from apathy. What is important to know is not whether the observations conform to the simplifying hypothesis, but whether they are compatible with seriously different hypotheses that are equally probable a priori. The correct procedure to follow, in order to decide whether the simplifying hypothesis should be made, seems to be the following. We first examine all available data in various ways, no doubt calculating the values of various test criteria, in order to form a judgment as to what kinds of departure from the ideal conditions occur. Significance tests as such are not useful, but we shall probably wish to have some idea of the possible sampling variation of our statistics. We then try to formulate a plausible class of admissible hypotheses, introducing as few extra parameters as possible. If we are lucky, we may feel we can get away with only one extra parameter. Let us consider specially this possibility. An instance would occur if we decided that the ideal statistical conditions held very closely provided we replaced "normal distribution" by "Pearson Type VII distribution"; there would then be one extra shape parameter, the exponent. Another instance would

occur if we decided that the ideal conditions held very closely except for dependence of the error variance on the mean, as defined in section 3.1; χ would be the one extra parameter.

Let us call the extra parameter δ, so scaled that when $\delta = 0$ we have the full ideal conditions. We must now decide how we should proceed if we knew for sure that δ was substantially different from 0. The answer would depend on the class of admissible hypotheses, that is, on what δ represented. It might be one of the following: (a) transform the observations and then use ordinary least squares, (b) use some kind of weighted least squares, with weights depending on the residuals and therefore determined iteratively, (c) apply the least-squares method to a nonlinear hypothesis about (μ_i), (d) abandon a comprehensive analysis of the observations and attempt only a more limited piecemeal analysis. Presumably this procedure would be less attractive than the ordinary least-squares analysis would have been, had we known for sure that δ was zero, because of greater computational effort, or because the parameters would be more poorly estimated, or because the results would be more difficult to state and comprehend, or because the results would be more modest in scope. If, however, the least-squares method were used when δ was not zero, the results would be to some extent in error and misleading. What we must now do is determine, as well as we can, the "break-even point," determine how far δ must be from zero for the error in using simple least squares to outweigh the disadvantages attending the alternative procedure. (There may be two break-even points, one positive, the other negative, but for brevity we shall speak as though there was one.) Once the break-even point is fixed, it is easy to formulate a well-defined decision problem. No doubt our prior opinion about the value of δ is diffuse, and some suitable probability distribution, possibly uniform, can be named, it matters little what; and some reasonable loss function, possibly quadratic, can be named—again it matters little what, provided the break-even point is observed. If the total sample information available about δ gives us a rather precise estimate of δ, then an almost optimum decision rule is to decide in favor of simple least squares or the alternative procedure according to which side of the break-even point the estimate of δ comes (see [2]).

One component of the above train of argument has received some attention in the literature, namely, to determine how much the results of a simple least-squares analysis are invalidated when δ differs from zero. Unfortunately, attention has been paid exclusively to the significance levels of certain tests concerning (θ_r). In most circumstances such tests are inappropriate and ought not to be made.

6.5. To sum up: In sections 2 and 3 we have considered four statistics designed to reveal certain types of departure from the ideal statistical conditions. Information has been given about their sampling distribution under the "null hypothesis" of the full ideal conditions. That is better than no information at all about sampling distributions, and can be directly applied to (approximate) significance tests having merit as complacency removers. Thus a modest contribution has

been made. A thorough investigation of the appropriateness of the least-squares method would have to go further, and would encounter grave difficulties. I suppose that no convincing investigation of this sort has ever been made, for any field of observation.

As for outliers, significance tests are only relevant if the question at issue is whether extreme observations, suggestive of gross errors or blunders, occur with a frequency incompatible with the ideal conditions. Such a question can well be asked when a considerable bulk of observations of a certain sort are being reviewed. For most fields of observation, one may expect that the answer will turn out to be yes. The day-to-day problem about outliers is different from this, however. It is not: is the ordinary least-squares method appropriate? but: how should the ordinary least-squares method be modified? not: do gross errors occur sometimes? but: how can we protect ourselves from the gross errors that no doubt occasionally occur? The type of insurance usually adopted (it is not the only kind conceivable) is to reject completely any observation whose residual exceeds a tolerance calculated according to some rule, and then apply the least-squares method to the remaining observations. In [3] suggestions were made for choosing a routine rejection rule, based on no more prior knowledge about gross errors than a belief that they occur sometimes. De Finetti [11] has considered a fully Bayesian approach to the rejection of outliers, necessarily based on more definite prior knowledge.

REFERENCES

[1] F. J. ANSCOMBE, Contribution to discussion of paper by G. E. P. Box and S. L. Andersen, *J. Roy. Statist. Soc., Ser. B*, Vol. 17 (1955), pp. 29–30.

[2] ———, Contribution to discussion of paper by F. N. David and N. L. Johnson, *J. Roy. Statist. Soc., Ser. B*, Vol. 18 (1956), pp. 24–27.

[3] ———, "Rejection of outliers," *Technometrics*, Vol. 2 (1960), pp. 123–147.

[4] F. J. ANSCOMBE and J. W. TUKEY, "The criticism of transformations" (abstract), *J. Amer. Statist. Assoc.*, Vol. 50 (1955), p. 566.

[5] M. S. BARTLETT, "Some notes on insecticide tests in the laboratory and in the field," *J. Roy. Statist. Soc.*, Vol. 3 Suppl. (1936), pp. 185–194.

[6] C. I. BLISS and A. R. G. OWEN, "Negative binomial distributions with a common k," *Biometrika*, Vol. 45 (1958), pp. 37–58.

[7] G. E. P. BOX and J. S. HUNTER, "Multi-factor experimental designs for exploring response surfaces," *Ann. Math. Statist.*, Vol. 28 (1957), pp. 195–241.

[8] A. S. C. EHRENBERG, "The unbiased estimation of heterogeneous error variances," *Biometrika*, Vol. 37 (1950), pp. 347–357.

[9] D. A. EVANS, "Experimental evidence concerning contagious distributions in ecology," *Biometrika*, Vol. 40 (1953), pp. 186–211.

[10] T. S. FERGUSON, "On the rejection of outliers," *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley and Los Angeles, University of California Press, 1961, Vol. 1, pp. 253–287.

[11] B. DE FINETTI, "The Bayesian approach to the rejection of outliers," *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley and Los Angeles, University of California Press, 1961, Vol. 1, pp. 199–210.

[12] R. A. FISHER, "The moments of the distribution for normal samples of measures of de-

parture from normality," *Proc. Roy. Soc. London, Ser. A*, Vol. 130 (1930), pp. 17–28. (Reprinted in *Contributions to Mathematical Statistics*, New York, Wiley, 1950.)

[13] ———, *Statistical Methods for Research Workers*, Edinburgh and London, Oliver and Boyd, 1932 (4th ed.), appendix to chapter III.

[14] F. GRAYBILL, "Variance heterogeneity in a randomized block design," *Biometrics*, Vol. 10 (1954), pp. 516–520.

[15] F. E. GRUBBS, "On estimating precision of measuring instruments and product variability," *J. Amer. Statist. Assoc.*, Vol. 43 (1948), pp. 243–264.

[16] P. G. MOORE and J. W. TUKEY, Answer to query 112, *Biometrics*, Vol. 10 (1954), pp. 562–568.

[17] M. E. TERRY, "On the analysis of planned experiments," *National Convention Transactions 1955, American Society for Quality Control*, pp. 553–556.

[18] J. W. TUKEY, "One degree of freedom for nonadditivity," *Biometrics*, Vol. 5 (1949), pp. 232–242.

[19] ———, Answer to query 113, *Biometrics*, Vol. 11 (1955), pp. 111–113.

[20] ———, "On the comparative anatomy of transformations," *Ann. Math. Statist.*, Vol. 28 (1957), pp. 602–632.

[21] F. YATES, *The Design and Analysis of Factorial Experiments*, Harpenden, England, Imperial Bureau of Soil Science (Technical Communication No. 35), 1937, section 10.