Biometrika (1983), **70**, 2, *pp.* 421–31
Printed in Great Britain

Robust statistical analysis of interlaboratory studies

BY DAVID M. ROCKE
Graduate School of Administration, University of California, Davis, California, U.S.A.

SUMMARY

A common procedure in testing analytical methods is to send a portion of each of a number of samples to each of several laboratories. The results of such a study are submitted to statistical analysis to determine the two important variance components in the problem: replication error and laboratory bias. Outliers are relatively common in these data both among laboratory effects and among the residuals. This paper presents a method of analysis for interlaboratory studies that is robust to the existence of outliers and long-tailed distributions of random effects. Theoretical considerations as well as a Monte Carlo study are adduced as support for this new technique.

Some key words: Biweight estimate; Huber estimate; Least squares; Monte Carlo method; Outlier; Random effects; Robust estimation; Variance component.

1. INTRODUCTION

Although much work has been done in the past 15 years on robust statistical methods, most researchers have concentrated on fixed effects models. Research on estimation of variance-like quantities has largely been restricted to auxiliary scale estimates for estimating the uncertainty of coefficients and to covariance and correlation matrices (Huber, 1981; Gnanadesikan, 1977). In part, this is because random effects models play a smaller part in statistical practice than fixed effects models, but measures of variation are also conceptually less well defined than measures of location. It may be difficult in many cases even to know what one should be estimating, much less by what method to do so. There is some research on the robustness of the F test and on robust estimation and testing for variance components (Arvesen & Layard, 1975; Atiqullah, 1962a, b; Tan & Wong, 1980; Tiao & Ali, 1971). For related work for the single-sample problem, see Healy (1978, 1979).

To analyse a random effects problem it is first necessary to understand the use to which this analysis will be put. Since this often depends on the context within which the problem occurs, the next section of this paper describes the food and agricultural chemistry application that motivated the work. The following section discusses general problems in defining variance-like quantities that satisfy properties often held to be desirable. The proposed new method of analysis is then introduced, its large sample properties are discussed, and small sample characteristics are tested by a Monte Carlo study. Finally, an example is presented of the use of this method on data arising from the applications context.

2. INTERLABORATORY STUDIES

Analytical tests of food and agricultural products are often required for quality control and regulatory purposes. One may need to know the percentage of protein in animal feed, the extent of contamination of tuna fish by mercury, the amount of

aflatoxin in peanut butter, or whether a macaroni product contains rat hairs. In order for the results of such tests to be meaningful, procedures must be well developed enough that a reanalysis does not drastically change the conclusions, and well enough specified that different laboratories will achieve similar conclusions from the same sample. Most such procedures are tested before adoption by a collaborative interlaboratory study to determine replication error and the extent of laboratory bias. In a typical such study, several samples to be analysed are divided, and a part of each sample is sent to each of a number of laboratories. The resulting data are analysed by the referee to yield variance component estimates for replication error and for laboratory bias.

Table 1 shows the data from a study on determining the amount of nicotine in a sample (Wagner & Thaggard, 1979). These data were originally analysed by a method

Table 1. *Data from Wagner & Thaggard* (1979)

Laboratory	Sample 1	2	3	4	5	6	7	8	9	10
A	0·161	0·192	0·373	0·417	0·663	0·692	0·952	0·918	1·288	1·186
B	0·157	0·193	0·366	0·412	0·660	0·688	0·965	0·933	1·260	1·179
C	0·159	0·193	0·373	0·406	0·643	0·671	0·954	0·905	1·193	1·146
D	0·294*	0·206*	0·405	0·429	0·630	0·701	0·982	0·964	1·213	1·208
E	0·152	0·188	0·374	0·419	0·681	0·714	0·967	0·950	1·243	1·182
F	0·123	0·157	0·331	0·374	0·621	0·662	0·949	0·879	—	—
G	0·152	0·176	0·350	0·390	0·657	0·652	0·947	0·877	1·218	1·147
H	0·163	0·196	0·348	0·416	0·637	0·674	0·959	0·898	1·198	1·154
I	0·183	0·217	0·391	0·426	0·685	0·675	0·946	0·930	1·253	1·197
J	0·164	0·165	0·369	0·406	0·658	0·670	0·944	0·915	1·196	1·147
K	0·151	0·186	0·375	0·406	0·658	0·696	0·961	0·965	1·219	1·198
L	0·146	0·180	0·360	0·421	0·661	0·668	0·976	0·986	1·350*	1·291*
M	0·160	0·195	0·372	0·413	0·651	0·682	1·005	0·908	1·206	1·152
N	0·110	0·136	0·237	0·252	0·520	0·540	0·814	0·759	1·020	0·947

Data are amounts of nicotine in mg extracted from Cambridge filter pads. Determinations are by gas–liquid chromatography. Values marked * were rejected as outliers in original analysis using Dixon's (1953) criterion. Laboratory N was rejected by Youden's (1975) test.

due to Youden (1975) that divides the samples into similar pairs, but one could also use standard analysis of variance techniques to estimate the variance components (Searle, 1971, § 10·2). Suppose one posits the model

$$y_{ij} = m + a_i + b_j + e_{ij},$$

where the a_i are the sample effects, constrained to sum to 0, the b_j are the random laboratory effects assumed to have mean 0 and variance σ_L^2, and the e_{ij} are the replication errors assumed to have mean 0 and variance σ_E^2. Then the F test for the existence of a laboratory effect has $F(13, 115) = 22{\cdot}76$ ($p < 0{\cdot}0001$), so that one is essentially certain of the existence of a laboratory effect. The replication standard deviation calculated by this method is 0·0277 and the laboratory bias standard deviation is 0·0410. The results reported in the original article are substantially different from these values. There are five numbers reported for each of the standard deviations, because of the pairwise analysis, and the root mean square averages of these are 0·015 for the replication error and 0·015 for the laboratory bias.

The explanation for this disparity lies in the rejection by the original authors of all the measurements from one laboratory and of four other data values as outliers. This is presumably because of a desire that unusual laboratories and individually aberrant

values should not be included in the analysis of a collaborative study to avoid inflation of the estimates of replication standard deviation and laboratory bias. In fact, the use of a laboratory rejecting procedure (Youden, 1975) and a data-value rejecting procedure (Youden, 1975; Dixon, 1953) has become routine. Nevertheless, such methods may induce a certain uneasiness in a statistician due to a problem of conflicting objectives. As Youden (1975, p. 30) puts it:

> If . . . the results come from different laboratories, it hardly makes sense to discard a fair proportion of the population of laboratories. These are the only laboratories we have and, anyway, we have no power to make them vanish. Our task is that of presenting a realistic picture of the population of laboratories. This last objective has to be balanced against the distortion of the picture that would occur from keeping a result so out of line that the estimate of error does not mirror the real merit of the analytical method.

3. Measures of variation for random effects models

One solution to the dilemma outlined in the last section is to employ a measure of variation other than the variance. In some sense, the two-stage procedure of outlier rejection and least squares analysis that is now used does that, but it does so in an ill-defined manner. The search for alternatives is somewhat constrained by the following result.

Theorem. *Let V be a functional defined on distribution functions F with the convention that, if X has distribution F, then $V(X)$ means $V(F)$. Suppose V has the following properties:*

(i) *if V is defined for X, then V is defined for $aX+b$ and $V(aX+B) = a^2\,V(X)$;*

(ii) *if V is defined for X and Y, where X and Y are independent, then V is defined for $X+Y$ and $V(X+Y) = V(X)+V(Y)$.*

This last property is needed to give the term variance components a meaning. Then V is a constant multiple of the variance functional whenever both are defined.

Proof. Let $X_1, X_2, \ldots$ be independent, identically distributed random variables having mean 0 and variance σ^2. Suppose $V(X_j) = v$ and let $Y_n = (X_1 + \ldots + X_n)\,n^{-\frac{1}{2}}$. Repeated use of (i) and (ii) shows that $V(Y_n) = v$ also. Since X_j has finite variance, Y_n converges to a normal variate with mean 0 and variance σ^2. In the limit, this implies that V takes the same value on X_j as on a normally distributed random variable with the same mean and variance. By (i) this is $K\sigma^2$ where $K = V(\Phi)$.

It is thus necessary to give up something in the definition of a variance-like functional to use in the analysis of random effects models, and the choice made here is to cede additivity, property (ii). Hence, if one speaks hereafter of a new method of estimating variance components, neither the 'variance' nor the 'components' is strictly correct. Nevertheless, it remains a useful locution for the method proposed in this paper.

Once it is agreed that other measures of variation than the variance may be considered, a difficult problem of choice presents itself. There is such a multiplicity of functionals that meet requirement (i) above that some principle of selection needs to be adopted.

There are several properties that logic and common usage dictate for such estimates. First, when random effects are normally distributed, the method should produce results very similar to the ordinary least squares analysis of the same data. This principle is violated by methods of outlier rejection. If, for example, one employs Youden tests on

the results of a collaborative study with normally distributed random effects, a laboratory will be eliminated from one study in 20. If a 5% sample-wise error rate is adopted for the Dixon test, as it usually is, then out of every 20 collaborative studies with 10 samples, about 12 will contain rejected observations. This preferential elimination of observations with large residuals implies that variance components from such studies will be biased downward and therefore that analytical methods may be thought to be more precise than they really are, an undesirable state of affairs.

Secondly, data that are in serious discord with the great majority of the results should be removed or downweighted in the analysis. Common practice is that laboratories that are too different from the remainder should not be used and that any value of a particular sample that is too discrepant should also be removed (Youden, 1975). Among other things, this implies that a standard least squares analysis of the original data will not produce acceptable results.

4. ROBUST ESTIMATION OF RANDOM EFFECTS

One rule for judging variance estimation methods is that derived from thinking of the data as contaminated normal. Suppose, after transformation if necessary, that errors of analysis have a distribution F with

$$F = (1-p)\Phi(0, \sigma^2) + p\Phi(0, k\sigma^2),$$

where Φ is the normal distribution function and $k > 1$. The data are thought of as consisting of mostly 'good' values with a certain variance and occasional 'bad' values with a larger variance. One could then judge the quality of a variance-like functional V by how close $V(F)$ is to σ^2, the variance of the 'good' data. This point of view is probably widely held (Healy, 1978, 1979) and has many advantages. It is conceptually simple and corresponds to intuitive feelings of many for how robust variance estimation should work.

In this paper a different approach is used. The starting point of the argument is an analysis of what makes the variance a useful measure of variation. Essentially, this is due to the fact that, for least squares estimation under broad conditions, the variance of a parameter estimate is an easily calculated function of known quantities and the population error variance. Thus, unless one assumes normality, the usefulness of the sample variance is arguably dependent on the use of a least squares methods of analysis.

There is now, however, much evidence that least squares methods do not provide the best analysis of real, not necessarily normal data (Stigler, 1977; Rocke, Downs & Rocke, 1982). Suppose that it is decided to summarize replicate observations by a location estimator T which performs well with normal data and is robust to long-tailed departures from normality. A proper choice of T can have 95% efficiency or more under normality, while providing substantial benefit if the errors are long tailed (Andrews *et al.*, 1972). In this case, the variance of an error distribution F does not provide useful information to the user about the precision of summarized replicates: neither does the variance of the 'good' observations in a mixture of normals model. To define a variance-like quantity that is of use in this situation, let $V_n^T(F)$ be the variance of T for independent, identically distributed samples of size n drawn from F and let

$$V^T(F) = \lim_{n\to\infty} n V_n^T(F).$$

This quantity then plays the same role in precision estimation for T as σ^2 does for the mean.

The class of estimates considered here are M-estimates (Huber, 1981) of two types. In both cases, parameter estimates are obtained by minimizing the sum of some function ρ of the residuals. This function is typically quadratic at zero but increases less rapidly than a quadratic for larger values of the argument. The particular estimates employed here are due to Huber (1964; 1973; 1981, Chapter 6) and Tukey (Beaton & Tukey, 1974; Mosteller & Tukey, 1977). Both estimates are most easily described by one of the algorithms used for estimation, which iteratively replaces the actual data with pseudo-data (Bickel, 1976; Huber, 1981, § 6.7). Upon convergence, each pseudoobservation $\tilde{y}$ is a weighted average of the original observation y and the fitted value $\hat{y}$. If $y-\hat{y}$ is small, then $\hat{y}$ is equal to or close to y. If $y-\hat{y}$ is large, more weight is given to the fitted value.

For the Huber estimate, choose a constant c. If $|y-\hat{y}| \leqslant cs$, where s is a robust scale estimate consistent for σ under normality, then $\tilde{y} = y$. If $y > \hat{y}+cs$, then $\tilde{y} = \hat{y}+cs$ and if $y < \hat{y}-cs$, then $\tilde{y} = \hat{y}-cs$. For location, this is Winsorization at $\pm cs$ (Dixon & Tukey, 1968). For $c = 2{\cdot}0$ we refer to this estimate as H_{20}. For Tukey's biweight estimate,

$$\tilde{y} = \hat{y}+\psi\left(\frac{y-\hat{y}}{cs}\right)cs.$$

Here $\psi(u) = u(1-u^2)^2$ if $|u| < 1$ and 0 otherwise, with s the median absolute residual and $c = 9{\cdot}0$. This estimate is called B_{90}. Under normality, the biweight omits entirely observations that are more than six estimated standard deviations from their fitted values and downweights observations inside these limits according to the size of the residual.

When H_{20} or B_{90} is chosen for T, the variance-like functional will be denoted by $V^H(F)$ or $V^B(F)$ respectively. We use $V^T(F)$ to refer to a variance-like functional defined by an unspecified robust estimator. Some properties of these functionals are as follows.

(a) $V^H(\Phi) = 1{\cdot}01$ and $V^B(\Phi) = 1{\cdot}02$. This means that a normal distribution will be assigned a quantity very slightly larger than the variance.

(b) If F is a long-tailed distribution, then $V^T(F) < \mathrm{var}\,(F)$. This reduction may be as much as several hundred per cent in practically relevant cases.

(c) Suppose X and Y are independent random variables with long-tailed distributions F and G, but with finite variances σ_X^2 and σ_Y^2. Then $V^T(X+Y) \neq V^T(X)+V^T(Y)$. Clearly, however,

$$V^T(X)+V^T(Y) < V^T(X+Y) < \sigma_X^2+\sigma_Y^2.$$

If we have data drawn from a model with p parameters and it is desired to estimate $V^T(F)$, where F is the distribution function of the errors, we may do so using a formulation based on Hampel's (1974) influence curve with Huber's (1973) correction factor as described by Huber (1981, formula 7·6·5). If the problem has several variance components, then each of them can be estimated by applying the procedure to the effects estimates for that variance component.

First, the procedure will be illustrated for the single sample case, as when one laboratory wishes to assess its own performance. Using the data from Healy (1979) as an example, we have that the standard deviation of the data is 13·1, Healy's method yields 11·0, the Huber estimate H_{20} gives 11·2 and the biweight B_{90} gives 11·1. One data value seems to be an outlier; if it is eliminated, the standard deviation of the 'cleaned' data is 10·0. Notice the Healy, Huber and biweight methods all yield similar values, but the standard deviation of all the data is considerably larger and the standard deviation of the cleaned data is considerably smaller.

How are these figures to be interpreted? The standard deviation of the cleaned data, although often used, is practically uninterpretable. As noted above, when such a procedure is employed on normal samples it gives estimates of variance that are biased downwards. The Healy, Huber and biweight methods are all useful as general measures of precision that depend on the central part of the error distribution. In normal samples, all give values that average out near to the standard deviation. The Huber and biweight scale estimates have an additional interpretation that increases their usefulness. If the Huber estimate of location is used as a summary of the data, its standard error is $11{\cdot}2/\sqrt{21} = 2{\cdot}44$ and similarly for the biweight. If new sets of data of size n were generated from the same source, one would expect the mean of the unaltered data to have a standard error near $13{\cdot}1/\sqrt{n}$ and the Huber estimate to have a standard error near $11{\cdot}2/\sqrt{n}$. Note that Healy's estimate, although it is derived from the sample after trimming one point from each end, is not meant for use in providing confidence limits on the corresponding trimmed mean. Rather, the goal is to estimate the standard deviation of the uncontaminated part of the error distribution.

Let us now specialize to the case at hand, the interlaboratory study. Suppose we have observations $\{y_{ij}\}$ for samples $i = 1, \ldots, q$ and laboratories $j = 1, \ldots, r$. The model used is $y_{ij} = m + a_i + b_j + e_{ij}$, where the a_i are the sample effects, assumed to sum to 0, the b_j are the laboratory effects, taken to be independent and identically distributed with mean 0, and the e_{ij} are the replication errors, assumed to be independent and identically distributed with mean 0. First, the estimates $\hat{m}$, $\hat{a}_i$ and $\hat{b}_j$ of the parameters and an auxilliary scale estimate s_E are produced using one of the methods described above. Then, the b_j are submitted to an analogous location estimator producing a summary $\tilde{b}$ and a scale estimate s_L.

Pseudoobservations $\tilde{y}_{ij}$ are defined by

$$\tilde{y}_{ij} = \hat{m} + \tilde{b} + \hat{a}_i + cs_L \psi\{(\hat{b}_j - \tilde{b})/(cs_L)\} K_L + cs_E \psi\{\hat{e}_{ij}/(cs_E)\} K_E,$$

where $\hat{e}_{ij} = y_{ij} - (\hat{m} + \hat{a}_i + \hat{b}_j)$, and where K_E is the correction factor associated with the main M-estimation of the parameters and K_L is the correction factor associated with the location M-estimate applied to the laboratory effects. When these pseudoobservations are submitted to a standard analysis of variance, the mean squares produced for error and laboratory are $V^T(F)$ and $qV^T(H)$, where F is the distribution of the errors and H is the distribution of the laboratory effect plus the average of q errors. Note that, because of the above theorem, the variance component for laboratories is not then a direct estimate of $V^T(G)$, G being the distribution function of the laboratory effects, but is instead determined by the reduction of $\hat{V}^T(F)$ from $\hat{V}^T(H)$, divided by q. In practice, however, there is typically very little difference between these quantities. This method of pseudoobservations may also be applied to problems with more than the two variance components found here. For example, if replicate observations are available for each cell, the random interaction effects may also be estimated and the results of summarizing these by an M-estimate may be included in the pseudoobservations. This procedure will be called the Huber procedure or the biweight procedure if the $V^H(F)$ or $V^B(F)$ is used. The standard procedure will be called least squares.

Asymptotically, as q and r both increase, the estimates of the error and laboratory variance components clearly tend to $V^T(F)$ and $V^T(G)$; however, the small sample properties of this procedure are less clear. Generally, one can expect that the procedure will be better behaved the larger c is, since the method is then closer to least squares. If it were the case that values of c small enough to be useful in trimming outliers required

unreasonably large numbers of laboratories or samples, then the suggested procedure would be essentially useless. The subject of the next section is a Monte Carlo study that was performed to explore that question.

5. Small-sample properties

Table 2 shows the results of estimating the error variance in six situations for a Monte Carlo experiment of 500 repetitions. Some conclusions that can be drawn from this table

Table 2. *Monte Carlo estimates of error variance component by standard and robust methods*

Run	k	Error distr.	Error var. comp. σ^2	V^H	V^B	Error var. comp. estimates S^2	$\hat{V}^H$	$\hat{V}^B$
1	5	N	1·00	1·01	1·02	0·99 (0·36)	0·99 (0·36)	1·01 (0·37)
2	10	N	1·00	1·01	1·02	1·00 (0·15)	1·01 (0·16)	1·02 (0·16)
3	5	LT	1·90	1·37	1·30	1·84 (1·09)	1·69 (0·97)	1·62 (0·93)
4	10	LT	1·90	1·37	1·30	1·93 (0·59)	1·51 (0·35)	1·43 (0·33)
5	5	VLT	5·95	1·29	1·12	5·70 (7·86)	2·85 (4·28)	2·21 (3·48)
6	10	VLT	5·95	1·29	1·12	6·17 (4·06)	1·49 (0·38)	1·19 (0·25)

Results based on 500 replications; simulation is for $k \times k$ table; parenthetical entries are standard deviations across replications. Here N is standard normal; LT is a mixture of 90% standard normal and 10% normal with mean 0 and variance 10; VLT is a mixture of 95% standard normal and 5% normal with mean 0 and variance 100; S^2 is residual variance from two-way model; V^H is robust method using Huber's ψ and $c = 2{\cdot}0$; V^B is robust method using the biweight ψ with $c = 9{\cdot}0$.

are as follows.

(i) With normal errors, even for small tables, the robust Huber and biweight methods yield results essentially identical to least squares. This is true both for average value and for spread.

(ii) The Huber procedure is conservative in that it does not underestimate the quantity $V^H(F)$. It reduces the variance component estimate substantially below the least square value in long-tailed cases. Also, the variance of the Huber variance component estimate is lower than that of the least squares estimate for long-tailed errors.

Other studies (Rocke & Downs, 1981) have shown that $V^H(F)$ and $V^B(F)$ can be estimated very accurately in single samples of this size under similar error distributions. The present study demonstrates how much more difficult this estimation problem is for more complex models.

Table 3 shows the results of estimating the treatment variance component by the three methods. The results have been scaled to be comparable to the values in Table 2. The following conclusion can be drawn from the table.

(a) With normally distributed treatments, the Huber procedure gives results that average out virtually identical to least squares and the variance of the robust estimate is somewhat smaller. Note that the type of error distribution has little effect on the treatment variance component estimate.

(b) The biweight treatment estimate is very badly biased, especially in the 5 by 5 cases. The effect here is so strong as to virtually disqualify the biweight from use with this method of variance component analysis.

(c) When the treatment distribution is long tailed, the Huber procedure does not reduce the treatment variance component estimate in the 5 by 5 cases. With only 5 treatment values to work from, effective reduction seems difficult. However, one is, at least, no worse off than with least squares.

(d) For the 10 by 10 cases with long-tailed treatments, the Huber estimate averages out lower than the least squares estimate, although the reduction is substantial only for the very long-tailed distributions.

Table 3. *Monte Carlo estimates of treatment variance component by standard and robust methods*

Run	k	Error distr.	Treatment distr.	Treatment var. comp. σ^2	V^H	V^B	Treatment var. comp. estimates S^2	$\hat{V}^H$	$\hat{V}^B$
1	5	N	N	1·00	1·01	1·02	0·95 (0·81)	0·92 (0·81)	0·55 (0·64)
2	10	N	N	1·00	1·01	1·02	1·00 (0·62)	1·01 (0·64)	0·80 (0·57)
3	5	LT	N	1·00	1·01	1·02	0·96 (0·83)	0·96 (0·82)	0·57 (0·65)
4	10	LT	N	1·00	1·01	1·02	1·01 (0·65)	1·02 (0·63)	0·80 (0·55)
5	5	VLT	N	1·00	1·01	1·02	0·96 (0·81)	0·97 (0·76)	0·57 (0·57)
6	10	VLT	N	1·00	1·01	1·02	1·01 (0·64)	1·02 (0·52)	0·80 (0·44)
7	5	N	LT	1·90	1·37	1·30	1·93 (2·85)	1·93 (2·84)	1·04 (1·90)
8	10	N	LT	1·90	1·37	1·30	1·82 (1·81)	1·72 (1·61)	1·33 (1·29)
9	5	LT	LT	1·90	1·37	1·30	1·92 (2·81)	1·92 (2·81)	1·06 (1·84)
10	10	LT	LT	1·90	1·37	1·30	1·85 (1·87)	1·75 (1·65)	1·33 (1·24)
11	5	VLT	VLT	5·95	1·29	1·12	5·82 (16·86)	5·84 (16·97)	3·10 (11·53)
12	10	VLT	VLT	5·95	1·29	1·12	5·78 (12·31)	3·48 (6·44)	2·26 (4·36)

For definitions of symbols, see footnote to Table 2. Error distribution is scaled in each case so that power of least squares F test for existence of treatment effect is about 0·7 to 0·8.

In summary, the Huber procedure produces estimates that are trustworthy over broad conditions. Even though the variance component estimates are not reduced as far as would be optimal, the reductions are substantial and useful. Compared to least squares, the Huber procedure is clearly superior. The biweight method, on the other hand, appears to be substantially flawed and not yet suitable for practical use.

Table 4. *Monte Carlo estimates of the probability of rejection of the hypothesis of no treatment effect using standard and robust tests*

Run	k	Treatment distr.	Error distr.	pr (rejection) P_L	P_H
1	5	N	N	0·73	0·73
2	10	N	N	0·85	0·84
3	5	N	LT	0·74	0·76
4	10	N	LT	0·84	0·89
5	5	N	VLT	0·80	0·90
6	10	N	VLT	0·84	0·99
13	5	0	N	0·052	0·052
14	10	0	N	0·040	0·050
15	5	0	LT	0·042	0·050
16	10	0	LT	0·048	0·046

All tests at nominal 0·05 level. Here P_L is probability of rejection using standard F test; P_H is probability of rejection using equivalent test run on pseudo-observations using Huber's ψ with $c = 2·0$. For definition of symbols, see footnote to Table 2.

Under some circumstances, it may be useful to perform a test for zero treatment variance. Table 4 gives the results of the standard F test and robust equivalent using the Huber procedure. As can be seen, both perform close to the nominal level and both have similar power when errors are normally distributed. In the presence of very long-tailed errors, the gain in power by using the Huber procedure can be very large.

6. An example

In this section the data from Table 1 are analysed by three methods. Before analysis, logs were taken because of an increase in the variance with the mean of the sample. The first method is the standard least squares analysis of the unaltered data. The missing values were iteratively replaced by fitted values until convergence, and the degrees of freedom for error correspondingly reduced by 2. The variance components were estimated by the analysis of variance method (Searle, 1971, § 10·2). The second is a least squares analysis of the 'cleaned' data; that is, of the data after removing those points that were eliminated in the original analysis. The data from laboratory N were removed entirely while the other values were treated as missing in the same manner as the two actual missing values. This will be referred to as the outlier removal method. The third method of analysis is the proposed new method using Huber's ψ and $c = 2{\cdot}0$. Table 5 gives a summary of these three analyses. The new method gives results that are intermediate between keeping all the data and removing suspected outliers entirely. Note the very substantial reductions in the variance component estimate achieved by the new method over least squares. The reduction in the laboratory variance is especially notable in view of the results in Table 3. The tails of the effects distributions in this case must be very long indeed.

Table 5. *Summary of results for analysis of logarithms of Table* 1 *data by three methods*

Method	F	$\hat{V}$(error) $\times 10^4$	$\hat{V}$(labs) $\times 10^5$
Least squares	14·88	50	169
Outlier removal	7·06	17	10
Huber procedure	8·38	19	14

Least squares is standard analysis of variance; outlier removal is a least squares analysis of data after removal of those data omitted in original paper; Huber procedure is method proposed in this paper using Huber ψ with $c = 2{\cdot}0$; $\hat{V}$(error) is error variance component estimate; $\hat{V}$(labs) is laboratory variance component estimate.

It is important to consider how the numbers in Table 5 are to be interpreted by the practising laboratory chemist. To some extent, any of the figures may be used as general-purpose nonspecific measures of variability. The least squares variance components, however, are unsatisfactory for practical use due to tremendous inflation by outliers; either the outlier removal method or the Huber procedure is usable for measuring variability of the central portions of the data. The Huber procedure has an important advantage over outlier removal in that it does not lead to downwardly biased variance estimates under normality. In this context, the Huber procedure may be thought of as an outlier downweighting technique that can be used instead of one of the common outlier rejection methods such as the Dixon test (Barnett & Lewis, 1978). In this case, the results from the Huber procedure can be interpreted in exactly the same way as those

from outlier removal except that extra protection is given from believing that analytical methods are more accurate than they really are.

The robust Huber method has another important advantage over outlier removal. Suppose that, later, one intends to take 10 replicate measurements and report a summary value. There is no reason to believe that the precision estimate derived from outlier removal will be useful here. The precision estimate derived from the Huber procedure which is $\sqrt{(0{\cdot}0019/10)} = 0{\cdot}014$, is a useful external estimate of the standard deviation of the H_{20} location estimator applied to the 10 new replicate values.

It may be thought to be undesirable that the robust variance component estimates depend on the choice of an estimation method. Yet, this is true of all the current methods as well. The least squares variance component estimates are arguably useful only so long as means and other least squares methods are to be used on unscreened future data, an unlikely possibility. Methods of cleaning data give values that depend on the choice of technique. If one does not wish to choose one method of future analysis, then values can be given for several possibilities, which increase the amount of useful information generated by the variance components analysis.

Suggestions from a referee which improved the clarity of the presentation are gratefully acknowledged, This research was supported by the Academic Senate of the University of California.

REFERENCES

ANDREWS, D. F., BICKEL, P. J., HAMPEL, F. R., HUBER, P. J., ROGERS, W. H. & TUKEY, J. W. (1972). *Robust Estimates of Location*. Princeton University Press.

ARVESEN, J. N. & LAYARD, M. W. J. (1975). Asymptotically robust tests in unbalanced variance components models. *Ann. Statist.* **3**, 1122–34.

ATIQULLAH, M. (1962a). On the effect of non-normality on the estimation of components of variance. *J. R. Statist. Soc.* B **24**, 140 7.

ATIQULLAH, M. (1962b). The estimation of residual variance in quadratically balanced least-squares problems and the robustness of the *F*-test. *Biometrika* **49**, 83–91.

BARNETT, V. & LEWIS, T. (1978). *Outliers in Statistical Data*. New York: Wiley.

BEATON, A. E. & TUKEY, J. W. (1974). The fitting of power series, meaning polynomials, illustrated on band-spectroscopic data. *Technometrics* **16**, 147 85.

BICKEL, P. J. (1976). Another look at robustness: A review of reviews and some new developments. *Scand. J. Statist.* **3**, 145 68.

DIXON, W. J. (1953). Processing data for outliers. *Biometrics* **9**, 74–89.

DIXON, W. J. & TUKEY, J. W. (1968). Approximate behavior of the distribution of Winsorized *t* (Trimming/Winsorization 2). *Technometrics* **10**, 83 98.

GNANADESIKAN, R. (1977). *Methods for Statistical Data Analysis of Multivariate Observations*. New York: Wiley.

HAMPEL, F. R. (1974). The influence curve and its role in robust estimation. *J. Am. Statist. Assoc.* **69**, 383 93.

HEALY, M. J. R. (1978). A mean difference estimator of standard deviation in symmetrically censored normal samples. *Biometrika* **65**, 643 6.

HEALY, M. J. R. (1979). Outliers in clinical chemistry quality-control schemes. *Clin. Chemistry* **25**, 675–7.

HUBER, P. J. (1964). Robust estimation of a location parameter. *Ann. Math. Statist.* **35**, 73 101.

HUBER, P. J. (1973). Robust regression: Asymptotics, conjectures, and Monte Carlo. *Ann. Statist.* **1**, 799–821.

HUBER, P. J. (1981). *Robust Statistics*. New York: Wiley.

MOSTELLER, F. & TUKEY, J. W. (1977). *Data Analysis and Regression*. Reading, Mass: Addison-Wesley.

ROCKE, D. M. & DOWNS, G. W. (1981). Estimating the variances of robust estimators of location: Influence curve, jackknife, and bootstrap. *Comm. Statist.* B**10**, 221 48.

ROCKE, D. M., DOWNS, G. W. & ROCKE, A. J. (1982). Are robust estimators really necessary? *Technometrics* **24**, 95 101.

SEARLE, S. R. (1971). *Linear Models*. New York: Wiley.
STIGLER, S. M. (1977). Do robust estimators work with real data? *Ann. Statist.* **5**, 1055–98.
TAN, W. Y. & WONG, S. P. (1980). On approximating the null and nonnull distributions of the F-ratio in unbalanced random-effect models from non-normal universes. *J. Am. Statist. Assoc.* **75**, 655–62.
TIAO, G. C. & ALI, M. M. (1971). Effect of non-normality on inferences about variance components. *Technometrics* **13**, 635–50.
WAGNER, J. R. & THAGGARD, N. A. (1979). Gas–liquid chromatographic determination of nicotine contained on Cambridge filter pads. *J. Assoc. Off. Anal. Chem.* **62**, 229–36.
YOUDEN, W. J. (1975). Statistical techniques for collaborative tests. In *Statistical Manual of the Association of Official Analytical Chemists*, pp. v–63. Arlington, Va: Assoc. of Off. Anal. Chem.

[*Received March* 1982. *Revised September* 1982]