JOURNAL OF RESEARCH of the National Bureau of Standards—B. Mathematics and Mathematical Physics
Vol. 70B, No. 2, April–June 1966

Treatment of Outliers in Samples of Size Three*

F. J. Anscombe[1] and Bruce A. Barron[2]

(February 10, 1966)

A reading that is a long way from most of the others in a series of replicate determinations is called an outlier. A particular procedure for rejecting outliers, and also a particular procedure for modifying outliers, are considered for samples of size three, supposed drawn from a common normal population except that one of the three readings may have an added bias. Numerical results are given illustrating the effects of the procedures on estimation of the location parameter. The calculations support a tentative general conclusion that estimation by least squares should usually be tempered by successive application of both a rejection rule and a modification rule.

Key Words: Statistics, outliers, residuals, estimation, robustness, least squares.

1. Introduction

In a situation where one or more parameters are to be estimated by the method of least squares, each observation is regarded as a given linear combination of the parameters plus a random observation error. When the parameters have been estimated, for every observed value a corresponding fitted value can be calculated; the difference between the observed and the fitted value is the residual. Any observed value for which the residual is much larger in magnitude than most of the other residuals is called an outlier.

It has often been suggested that outliers should be treated differently from other observations. Three ways of treating them are

(i) retain the outliers as they stand, giving all observations equal weight;

(ii) reject the outliers, which means giving them zero weight;

(iii) retain the outliers with reduced weight—this is equivalent to modifying the outliers so that they become less different from the other observed values, and then giving them full weight.

This study deals with the effect of using a particular rejection procedure, or alternatively a particular modification procedure, for outliers in samples of size three. The three readings are assumed to be random independent observations drawn from a normal parent population, except that one of the readings may have an added bias. Our object is to estimate the mean of the normal population.

To suppose that some readings are all drawn from a common normal population, except that one has an added bias, may seem an implausible way to represent reality, but consideration of this situation throws useful light on the outlier problem, at less computational expense than some other schemes.

A sample of size three is the smallest for which a study of rejection or modification of outliers, with generation of precise numerical results, is not trivial. The results are of direct interest because chemical determinations are often made in triplicate. Moreover, they provide a check on approximate results and conjectures relating to larger samples.

Lieblein [6] studied the effect of regularly discarding the most discrepant reading from a sample of size three and using the mean of the closest pair as estimate of the population mean. He found the variance of this estimate in various circumstances. Our investigation is similar in spirit to his, though there is no overlap in results.

Dixon [3] was perhaps the first to distinguish clearly two general problems concerning outliers: (a) the problem of identifying a "significant" outlier, in order to infer that something has gone wrong with the experimental procedure, or possibly to explore the outlier as an unusual occurrence of interest; and (b) the problem of obtaining a procedure of analysis not appreciably affected by the presence of abnormal observations. He pointed out that the second problem was important in the estimation of parameters in situations where unavoidable occasional contamination occurred. Using mean squared error as the basis of comparison, he examined several estimates of the population mean (sample mean, median, and mean after application of various rejection rules) under various assumptions of contaminated sampling. Samples of size 5 and 15 were considered, for which an attempt was made to formulate a recommended procedure for processing data for outliers.

Having unfortunately overlooked this work by Dixon, one of the present authors (Anscombe [1][3]) independently made more sweeping suggestions in the same direction: choice of an outlier rejection criterion could often appropriately be based on consideration of its effect on the mean squared error of estimates of the parameters of interest in a least squares analysis,

*Invited paper. Prepared in connection with research at Yale University supported by the Army, Navy, Air Force and NASA under a contract administered by the Office of Naval Research, task NR 042–242, contract Nonr 609(52).

[1] Yale University.
[2] Rockefeller University, New York 21, N.Y.

[3] Figures in brackets indicate the literature references at the end of this paper.

rather than on the traditional rate of rejection. The percentage increase in variance of estimation errors due to using the rule, when in fact all observations came from a homogeneous normal source, would be an appropriate measure of the cost or premium of the procedure; and the reduction in mean squared error when spurious readings were present would measure the protection given by the procedure.

Jeffreys [5] forcefully attacked the use of any outlier rejection rule on several grounds, one being that the resulting estimate of the population mean was a discontinuous function of the observations. He and others, notably Tukey [8] and Huber [4], have made suggestions for assigning reduced but not zero weight to outliers, the weight being a continuous function of the magnitude of the residual. The modification rule considered below is of Huber's type.

With these previous studies in mind, we now formulate procedures for treatment of outliers (in sec. 2) and consider their effectiveness (in secs. 3 and 4), for samples of any size. Then in section 5 our computations for samples of size three are presented and discussed. Tentative general conclusions are drawn in section 6. Some notes on the computations appear in section 7.

2. Definition of Estimation Procedures

Suppose we are given some observations $y_1, y_2, \ldots, y_n$, each of which is a determination or estimate of a common "true" value μ. We wish to combine the observations to form a single improved estimate of μ (or otherwise make inferences about μ).

It is convenient to define the sample mean $\bar{y}$, the residuals z_i $(i=1, 2, \ldots, n)$ and the number ν of residual degrees of freedom by

$$n\bar{y}=\sum_i y_i, \qquad z_i=y_i-\bar{y}, \qquad \nu=n-1. \tag{1}$$

The custom in this situation is to hope that (near enough) the y's are realizations of independent random variables each having the same normal distribution with mean μ.

If this hope were believed to be accurately fulfilled, the estimation problem would be well defined and easy. If the variance of the common normal distribution were supposed known, the sufficient statistics $\bar{y}$ and n would constitute a complete summary of the data; and if the variance were not known, the sufficient statistics $\bar{y}$, n, and $\Sigma_i z_i^2$ would be a complete summary. In either case, we could regard $\bar{y}$ as estimating μ, with the other statistics ancillary. As is well known, $\bar{y}$ is the value for μ at which the sum of squares

$$\Sigma_i(y_i-\mu)^2 \tag{2}$$

is minimized.

But ordinarily it is unreasonable to suppose that the hoped-for property of the y's is accurately true. If we think in terms of a single estimate of μ, $\bar{y}$ is not necessarily the best to choose. In particular, we should usually bear in mind that the observations may have a propensity towards outliers. Two alternative theoretical descriptions of an outlier phenomenon are (i) some of the observations are affected by a gross error or mistake, which adds a bias onto the reading that would otherwise be obtained, (ii) the distribution of deviations of the observations from μ is not normal, but has longer tails, like a logistic distribution, for example. (Further theoretical descriptions are easily invented.) We therefore consider how to define a function $\hat{\mu}$ of the observations that may possibly estimate μ satisfactorily when some kind of outlier phenomenon is present. The difficulty here arises from our reluctance to specify firmly the distribution of the observations in terms of a very few parameters.

The traditional way of treating outliers is to reject them according to some rule and then let $\hat{\mu}$ be the average of the remaining observations. We here consider the following rule, as an example. First a critical size K for a residual is chosen $(K>0)$. Then the rule is

Rejection rule. Let M be a value of i for which $|z_i|$ is greatest. (M may be expected to be unique if $n\geq 3$ and if the observations are recorded with high precision.) If $n=2$, or if $n\geq 3$ and also $|z_M|\leq K$, retain all observations and quote $\bar{y}$ as the estimate of μ. If $n\geq 3$ and also $|z_M|>K$, reject y_M from the sample and act as though the remaining $n-1$ observations were the whole sample. With the observations relabeled and n and z's redefined, go back to the start of this rule.

If the initial sample size is 2, the rule sets $\hat{\mu}$ equal to the simple mean $\bar{y}$ in any case. If the initial sample size is 3, the rule leads *either* to retention of all three observations with equal weight (if $|z_M|\leq K$), so that $\hat{\mu}=\bar{y}$, *or* to rejection of just one observation, so that $\hat{\mu}$ is the average of the other two observations. If the initial sample size n exceeds 3, the rule leads conceivably to rejection of any number of observations from 0 to $n-2$, inclusive.

To implement the suggestion that outliers ought to be given reduced but not zero weight, we also consider the following rule, which seems to be computationally the simplest possible such rule. A critical size K for a residual is chosen $(K>0)$. Then the rule is

Modification rule. Choose as $\hat{\mu}$ a value for μ at which

$$\Sigma_{(1)}(y_i-\mu)^2+\Sigma_{(2)}K\{2|y_i-\mu|-K\} \tag{3}$$

is minimized, where $\Sigma_{(1)}$ means summation over those values of i for which $|y_i-\mu|\leq K$ and $\Sigma_{(2)}$ means summation over the remaining values of i. ($\hat{\mu}$ is unique provided there is at least one value of i for which $|y_i-\hat{\mu}|<K$.)

This rule may be alternatively expressed by saying that $\hat{\mu}$ is chosen to minimize the sum of squares (2),

but under a condition that some observations are modified if necessary so that no residual exceeds K in magnitude. Specifically, each observation y_i such that $|y_i-\hat{\mu}|>K$ is modified to a value y_i^* such that $y_i^*-\hat{\mu}$ has the same sign as $y_i-\hat{\mu}$ but $|y_i^*-\hat{\mu}|=K$. Which observations are to be modified (or in the previous language, which observations are to be included in the second summation $\Sigma_{(2)}$) must in general be discovered in several steps of trial and error, a task of quadratic programming. (It has been considered, for general regression analysis, by Sand [7].)

When $n=2$, we may always set $\hat{\mu}=\bar{y}$. This is the unique possibility for $\hat{\mu}$ if the two observations are spaced not more than $2K$ apart. Otherwise $\hat{\mu}$ may be chosen anywhere in an interval of width $|y_1-y_2|-2K$, centered at $\bar{y}$.

When $n=3$, let $y_{(1)}$, $y_{(2)}$, $y_{(3)}$ denote the three observations, rearranged in ascending order of magnitude. Then $\hat{\mu}$ is determined as follows:

(i) If $\bar{y}-y_{(1)}$ and $y_{(3)}-\bar{y}$ are both not greater than K, no observation is modified and $\hat{\mu}=\bar{y}$.

(ii) If $y_{(2)}-y_{(1)}$ and $y_{(3)}-y_{(2)}$ both exceed K, $y_{(1)}$ and $y_{(3)}$ are both modified and $\hat{\mu}=y_{(2)}$, the median.

(iii) Otherwise, either $y_{(1)}$ or $y_{(3)}$, but not both, is modified. If, for example, $y_{(3)}$ is modified (because $y_{(3)}-\bar{y}>K$ and $y_{(3)}-y_{(2)}>K>y_{(2)}-y_{(1)}$), $\hat{\mu}$ is defined by

$$2\hat{\mu}=y_{(1)}+y_{(2)}+K.$$

$\hat{\mu}$ lies between the mean $\bar{y}$ and the median $y_{(2)}$. (In case (ii), the modified observations are $y_{(1)}^*=y_{(2)}-K$, $y_{(3)}^*=y_{(2)}+K$. In case (iii), if $y_{(3)}$ is modified, $y_{(3)}^*=\hat{\mu}+K$ and $3\hat{\mu}=y_{(1)}+y_{(2)}+y_{(3)}^*$.)

3. Distribution Assumption

The above rules yield estimates $\hat{\mu}$ of μ that are desensitized to outliers and may therefore be preferred to $\bar{y}$. For each rule, $\hat{\mu}$ is a function of the n observations and of K. We could distinguish the two functions with a suffix, but that will be unnecessary because we shall always make clear which rule is under discussion.

In order to assess the effectiveness of the rules, seeing how the effectiveness of each varies with K and how one rule compares with the other, we need to specify the true statistical properties of the observations. In this paper we suppose that the observations have the hoped-for property exactly, except that possibly *one* observation has an added bias. That is, we make the

Assumption: the y's are realizations of independent random variables each normally distributed with the same variance σ^2 and with these expectations:

$$E(y_i)=\mu\ (i=1,2,\ldots,n-1), \qquad E(y_n)=\mu+b\sigma. \tag{4}$$

We here take the liberty of using y_i both as the name of the random variable corresponding to the ith observation and as the name of that observation.

We shall be interested in the ratio of K to σ, which we denote by C, thus:

$$K=C\sigma. \tag{5}$$

We shall take expectations with respect to the above random variables, for fixed μ, σ, C, and b. Having σ and C fixed implies that K is fixed and therefore not determined by the observations themselves (as it would be, for example, if K were chosen to be equal to a given multiple of the sample standard deviation).

An alternative distribution assumption that would be interesting, but is not considered in this paper, would be that the observations were independently drawn from a common nonnormal distribution with mean μ, such as a logistic distribution.

4. Formulas

For better or worse, we shall assess the effectiveness of the two rules for treating outliers through the mean squared error of the sampling distribution of $\hat{\mu}$, under the above distribution assumption. Hopefully, when $b=0$, we shall find that the variance (which is also the mean squared error) of $\hat{\mu}$ is little larger than that of $\bar{y}$, namely σ^2/n, but when b is large $E(\hat{\mu}-\mu)^2$ will be smaller than

$$E(\bar{y}-\mu)^2=(\sigma^2/n)\{1+(b^2/n)\}. \tag{6}$$

To determine $E(\hat{\mu}-\mu)^2$ precisely, under the above distribution assumption, is a formidable task even when n is as small as 3. If we define an orthogonal linear transformation (Helmert transformation) of the y's, so that the new variables are $\bar{y}\sqrt{n}$, x_1, x_2, . . ., x_{n-1}, say, these variables are independently normally distributed with the same variance σ^2. We may write

$$\hat{\mu}-\mu=\{\bar{y}-\mu-(b\sigma/n)\}+U\sigma, \tag{7}$$

where U is a function of the x's but not of $\bar{y}$. Hence

$$E(\hat{\mu}-\mu)^2=(\sigma^2/n)+\sigma^2E(U^2),$$

or

$$(n/\sigma^2)E(\hat{\mu}-\mu)^2=1+nE(U^2). \tag{8}$$

$E(U^2)$ can be expressed as an $(n-1)$-dimensional integral. The form of the integrand depends on which of various linear inequalities among the x's are satisfied, so that the region of integration is divided into many zones, in each of which the integrand has a simple expression, different from zone to zone.

The main purpose of this paper is to present results of the calculation of $E(\hat{\mu}-\mu)^2$ when $n=3$. Values of $(3/\sigma^2)E(\hat{\mu}-\mu)^2$ are shown for an assortment of values

of C and b in table 1 (for the rejection rule) and table 2 (modification rule).

The case $n=3$ is important in its own right, and moreover may give some indication of the behavior of $E(\hat{\mu}-\mu)^2$ for larger values of n. For the latter purpose, it is useful to have some information about limiting values and approximations. These are now summarized.

Rejection rule. When $b=0$, the following approximate expression for the variance (mean squared error) of μ has been given [1]:

$$(n/\sigma^2)\,\mathrm{var}(\hat{\mu}) \doteq 1+(n/\nu)\{2t_\alpha\varphi(t_\alpha)+\alpha\}, \qquad (9)$$

where

$$t_\alpha = C\sqrt{n/\nu}, \qquad \alpha = 2\Phi(-t_\alpha)$$

and the functions Φ and φ are defined thus:

$$\Phi(t)=\int_{-\infty}^{t}\varphi(u)du, \qquad \varphi(u)=e^{-u^2/2}/\sqrt{2\pi}.$$

This result is asymptotically correct as $C\to\infty$ with n and ν fixed, and may be expected to be fairly good if C is somewhat greater than 2.

A more easily calculated formula, alternative to (9) when $b=0$, can be obtained, similar to one given by Anscombe and Tukey [2]:

$$(n/\sigma^2)\,\mathrm{var}(\hat{\mu}) \doteq 1+(n/\nu)\Phi(-N), \qquad (10)$$

where N is defined in terms of the above t_α by

$$t_\alpha = 1.40+0.85\,N.$$

When $b\to\infty$ with C fixed, rejection of the "bad" observation y_n becomes certain, and if no further rejection occurred we should have for the variance (mean squared error) of $\hat{\mu}$

$$(n/\sigma^2)\,\mathrm{var}(\hat{\mu}) \to 1+(1/\nu). \qquad (11)$$

This is accurately true when $n=3$, because no further rejections are allowed. For $n\geq 4$, the right side of (11) should be increased to allow for the effect of possible further rejections. A lower bound for $E(\hat{\mu}-\mu)^2$ when b is large but not infinite has been given ([1], eq (6.3)), but appears from the present calculations to be a very poor approximation.[4]

Modification rule. When $b=0$, the result for the modification rule corresponding to (9) above for the rejection rule is

$$(n/\sigma^2)\,\mathrm{var}(\hat{\mu}) \doteq 1+(n/\nu)\{(1+t_\alpha^2)\alpha-2t_\alpha\varphi(t_\alpha)\}, \qquad (12)$$

[4]The b-values in table 3 of [1] are presumably substantially too low, but possibly the remarks based on them are correct.

t_α and α being defined as before. This formula may be expected to be less helpful than (9). Both formulas are derived assuming C to be so large that few observations are rejected or modified (as the case may be) and the negative correlation between residuals within a sample is unimportant. But the values of C that are of practical interest are smaller for the modification rule than for the rejection rule.

When $b\to\infty$ with C fixed, modification of the "bad" observation y_n becomes certain, and if no further modification occurred we should have

$$(n/\sigma^2)E(\hat{\mu}-\mu)^2 \to 1+(1/\nu)+(nC^2/\nu^2). \qquad (13)$$

In fact, however, other observations may be modified, and because of the positive bias resulting from y_n the tendency will be for low readings to be modified upwards rather than for high readings to be modified downwards. It follows that the right side of (13) is too low. For $n=3$, $\nu=2$, it is not hard to show that the correct result is

$$(3/\sigma^2)E(\hat{\mu}-\mu)^2 \to \frac{3}{2}+\frac{3C^2}{4}+\frac{3C}{\sqrt{2}}\varphi\left(\frac{C}{\sqrt{2}}\right) +3\left(1-\frac{C^2}{2}\right)\Phi\left(-\frac{C}{\sqrt{2}}\right). \qquad (14)$$

Note. Although in section 2 we explicitly mentioned only the possibility that the y's were all determinations of the same location parameter μ, much of what has been said can be adapted to regression analysis, μ being replaced by a linear function of parameters β_r. Formulas (6), (9), (10), (11), (12), and (13) are valid with the following amendment and reinterpretation. The y's are observations in a factorial experiment with n experimental units (n being even) and an orthogonal design matrix. We focus attention on one particular two-level factor; this appears $n/2$ times at its upper level, $n/2$ times at its lower level. Let β_1 stand for one half of the response to this factor, so that $2\beta_1$ is the change in the expectation of an observation caused by changing the level of the factor from lower to upper; let b_1 denote the usual estimate of β_1 (total of observations for which the factor is at the upper level minus total of other observations, divided by n); and let $\hat{\beta}_1$ denote the estimate of β_1 yielded by the rejection or modification rule (as the case may be). Let $\{z_i\}$ be the residuals and ν the number of residual degrees of freedom after the estimation of all factor effects by least squares, and let the $n\times n$ matrix (q_{ij}) be defined by

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

We assume that for all i and j

$$q_{ii}=\nu/n, \qquad |q_{ij}|<\nu/n \qquad (i\neq j).$$

On the left sides of the above mentioned formulas, replace μ, $\bar{y}$, and $\hat{\mu}$ by β_1, b_1, and $\hat{\beta}_1$.

5. Results of Computation

Tables 1 and 2 present the mean squared error of $\hat{\mu}$ when $n=3$. The nine values for C are the "round" numbers 1, 1.5, 2, 3, and ∞, and also the values for which the entry in the first row (for $b=0$) is 1.04, 1.02, 1.01, and 1.005. The last column ($C=\infty$) refers to use of the unadjusted mean, $\bar{y}$, as the estimate of μ; the values are given by formula (6). Various values for the bias factor b are shown. In the last line of each table, except for the entry in the last column ($C=\infty$), the entries are exactly equal (to the number of decimal places shown) to the limiting values as $b\to\infty$ given in formulas (11) and (14).

The percentage increase in sampling variance resulting from use of $\hat{\mu}$ instead of $\bar{y}$, when all the observations have a common normal distribution (the hoped-for property), may be termed the premium charged by the rule. The premium may be read from the first line of tables 1 and 2 (for $b=0$) by subtracting the initial 1 and multiplying by 100.

In figure 1 mean squared error is graphed against b for (i) the unadjusted mean $\bar{y}$, (ii) $\hat{\mu}$ given by the rejection rule with 2 percent premium, (iii) $\hat{\mu}$ given by the modification rule, also with 2 percent premium.

Table 3 compares approximations (9), (10), and (12) with the correct values. (For (9) see also table 4 of [1].)

So far so good. The burning question that faces us is what do these computations show concerning the relative merits of the two methods of treating outliers, and if we decide to use either one, how should

TABLE 1. *Values of* $(3/\sigma^2)\mathrm{E}(\hat{\mu}-\mu)^2$ *for the rejection rule*

	$C=1.00$	1.50	2.00	2.46003	2.66184	2.84623	3.00	3.01724	∞
$b=0.00$	1.7318	1.4116	1.1496	1.0400	1.0200	1.0100	1.0054	1.0050	1.0000
.25	1.7731	1.4510	1.1812	1.0650	1.0433	1.0322	1.0270	1.0266	1.0208
.50	1.8931	1.5664	1.2753	1.1403	1.1133	1.0991	1.0921	1.0916	1.0833
1.00	2.3163	1.9896	1.6404	1.4422	1.3956	1.3688	1.3546	1.3534	1.3333
1.50	2.8488	2.5649	2.1969	1.9395	1.8682	1.8229	1.7969	1.7946	1.7500
2.00	3.3185	3.1297	2.8401	2.5914	2.5097	2.4522	2.4161	2.4126	2.3333
3.00	3.6602	3.6782	3.7634	3.8852	3.9340	3.9698	3.9917	3.9936	4.0000
4.00	3.2476	3.3066	3.5722	4.1362	4.4702	4.7967	5.0682	5.0980	6.3333
6.00	2.0323	2.0353	2.0674	2.2339	2.4143	2.6761	2.9854	3.0259	13.0000
8.00	1.5712	1.5712	1.5712	1.5737	1.5780	1.5872	1.6025	1.6048	22.3333
10.00	1.5045	1.5045	1.5045	1.5045	1.5045	1.5045	1.5045	1.5045	34.3333
12.00	1.5001	1.5001	1.5001	1.5001	1.5001	1.5001	1.5001	1.5001	49.0000
15.00	1.5000	1.5000	1.5000	1.5000	1.5000	1.5000	1.5000	1.5000	76.0000

TABLE 2. *Values of* $(3/\sigma^2)\mathrm{E}(\hat{\mu}-\mu)^2$ *for the modification rule*

	$C=1.00$	1.29420	1.50	1.52486	1.73307	1.92458	2.00	3.00	∞
$b=0.00$	1.0860	1.0400	1.0216	1.0200	1.0100	1.0050	1.0038	1.00004	1.0000
.25	1.1106	1.0631	1.0439	1.0422	1.0316	1.0263	1.0249	1.0209	1.0208
.50	1.1830	1.1315	1.1102	1.1082	1.0962	1.0899	1.0883	1.0834	1.0833
1.00	1.4512	1.3909	1.3651	1.3627	1.3480	1.3404	1.3385	1.3333	1.3333
1.50	1.8295	1.7743	1.7532	1.7516	1.7424	1.7400	1.7401	1.7487	1.7500
2.00	2.2380	2.2156	2.2189	2.2202	2.2354	2.2540	2.2616	2.3256	2.3333
3.00	2.8859	2.9970	3.1075	3.1223	3.2546	3.3830	3.4333	3.9014	4.0000
4.00	3.1789	3.4116	3.6342	3.6642	3.9389	4.2212	4.3376	5.7419	6.3333
6.00	3.2675	3.5619	3.8520	3.8920	4.2701	4.6861	4.8677	8.0425	13.0000
8.00	3.2687	3.5645	3.8566	3.8970	4.2789	4.7014	4.8867	8.3350	22.3333
10.00	3.2687	3.5646	3.8566	3.8970	4.2790	4.7014	4.8867	8.3396	34.3333

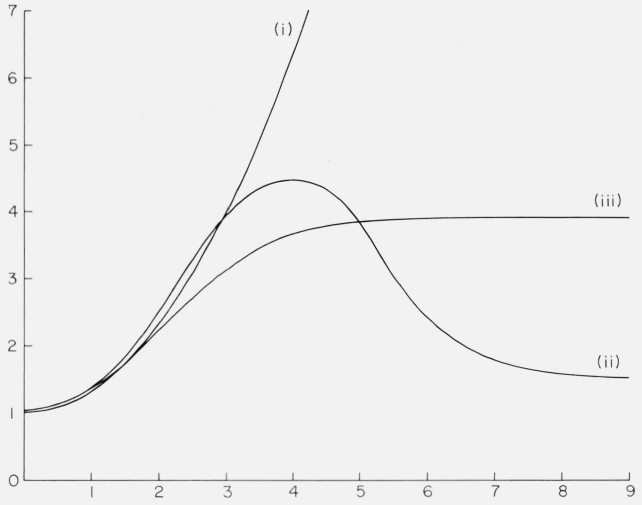

FIGURE 1. *Curves showing* $(3/\sigma^2)\mathrm{E}(\hat{\mu}-\mu)^2$ *as a function of* b.

K be chosen? No set of calculations such as ours can settle the matter beyond dispute, because we do not know that reality is well represented by the distribution assumption that we have used, nor by any other distribution assumption that we might have used instead. However, in the light of such theoretical knowledge as we have, the calculations do seem to support some clear-cut conclusions that may have general validity in regression analysis.

TABLE 3. *Approximations to* $(3/\sigma^2)$ *var* $(\hat{\mu})$ *when* b = 0

	$C=1.00$	1.50	2.00	3.00
Rejection rule:				
Correct Value	1.7318	1.4116	1.1496	1.0054
Approximation (9)	2.0234	1.5060	1.1674	1.0055
Approximation (10)	1.8725	1.4553	1.1627	1.0056
Modification rule:				
Correct value	1.0860	1.0216	1.0038	1.00004
Approximation (12)	1.1351	1.0277	1.0043	1.00004

We see from the tables that if b is small, in the range 0 to 1.5, both rules give estimates inferior to the unadjusted $\bar{y}$. If b is in a middle range, roughly from 2 to 4, the modification rule fares best. If b is above 5 or 6, the rejection rule is much better than the modification rule, which in turn is very much better than $\bar{y}$. These comparisons hold fairly consistently, when we compare a rejection rule with a modification rule *either* having the same premium (as in fig. 1) *or* having the same value for C. Thus which type of rule is to be preferred depends on how large we expect b will be (insofar as our distribution assumption can be accepted as a description of the facts).

Now we can distinguish two quite different processes that lead to outliers in a series of readings. On the one hand there may be mistakes or failures to do what

is intended—instrumental failures, errors in transcription or in arithmetic, mixed-up records, etc. A reading affected by an accident of this sort may easily lie a great distance from other readings. Therefore if we wish to guard against such gross errors, it is reasonable to choose a rejection rule. In practice, most observers discard extremely aberrant readings as obviously wrong, without any explicit rule, provided that they actually examine the readings. More and more nowadays the output of instruments is fed directly to a computer for processing, and then it is important that proper provision should be made for intercepting gross errors.

On the other hand, outliers may arise in good observations when no blunder or failure has occurred. The normal law of errors beloved of statistical theorists is not a law of nature, and observational errors do not have to conform to it. Probably many actual error distributions have somewhat longer tails than the normal. Jeffreys [5] reports some investigations of errors in astronomical readings, and suggests that a homogeneous series of readings by one observer may be expected to follow a Pearson Type VII distribution, having the same shape as a Student distribution with 7 deg of freedom. Tukey has pointed out (privately) that this distribution has nearly the same shape as a logistic distribution—the difference could hardly be detected empirically. In view of the distribution assumption of the present paper, it is interesting to note that the Student distribution with 7 deg of freedom is even more closely approximated by the distribution of the sum of two independent random variables, $X+Y$, where X is normally distributed and Y has chance 0.95 of being equal to 0 and chance 0.05 of being equal in magnitude to three times the standard deviation of X (positive or negative with equal chances).[5] This suggests that, to represent Jeffreys's type of long-tailed distribution of errors with the distribution assumption of this paper, we should regard b as taking the values 0 and 3 with something like a 6:1 frequency ratio. (Our distribution assumption cannot exactly represent samples of size 3 from the distribution of $X+Y$, because in a few such samples (less than 1 %) there would be more than one non-zero Y-value, and our calculations do not apply.) For these values of b, our tables indicate that the modification type of rule should be used.

This finding fits well with the consideration of maximum likelihood estimation of μ when the errors have a long-tailed distribution (see [2], sec. 8). In particular, maximum likelihood estimation of the location parameter of a logistic distribution is closely approximated by our modification rule, when K is about 1.1 times the true standard deviation of the logistic distribution, or about 1.25 times a pseudo standard deviation estimated from the slope of the middle part of the cumulative frequency curve plotted on "probability" graph paper (this being the sort of estimate of σ that we might make from past records if we believed that the error distribution was normal except for some outliers).

Thus our calculations support the following general conclusions which are closely in line with suggestions made by Tukey [8].

6. Tentative Conclusions

Whenever we think of applying the method of least squares to some readings in order to estimate a parameter or parameters of location, we shall do well to recognize the two possibilities that (i) occasionally a reading may be "bad", grossly in error and useless for the estimation purpose at hand (though possibly interesting for other reasons), and (ii) the "good" readings may have a somewhat longer-tailed distribution than the normal. In view of these possibilities (especially when the statistical analysis is computerized), it will be advisable to use first a rejection rule and then a modification rule. The rejection rule should have K so large that it will almost never reject "good" observations, but will protect against really "bad" ones. The modification rule will have a lower value for K and will aim to yield good estimates if the error distribution does not greatly differ from a normal or a logistic distribution. Actually the choice of K for the modification rule is likely to depend not only on considerations of efficiency of estimation but also on speed in computation. The smaller K is, the more iterations may be needed to carry out the modification procedure. It may therefore be wise to choose K so that not more than a few percent of readings (on the average) will be modified.

How to estimate from the data the precision of estimates obtained through the modification rule seems not to be well understood at present. But one thing at a time! (For a sample of size 3 such estimation is ludicrous anyway; hence the assumption in our calculations that σ was known.)

Of course when large collections of similar data are available for study, it is possible to investigate their statistical properties and adjust the estimation procedures accordingly. But in the absence of a special study it would be good routine practice always to temper the method of least squares by the combined rejection-modification procedure just outlined.

7. Notes on the Computation

We may set

$$x_1=\frac{y_1-y_2}{\sqrt{2}}=\frac{z_1-z_2}{\sqrt{2}}$$

$$x_2=\frac{2y_3-y_1-y_2}{\sqrt{6}}=z_3\sqrt{\frac{3}{2}}.$$

[5] The middle ordinate of the density function, multiplied by the standard deviation, makes a good index of shape. For (i) the Student distribution, (ii) the logistic, (iii) the distribution of $X+Y$, as specified, this index is approximately (i) 0.4555, (ii) 0.4534, (iii) 0.4566. Fisher's shape coefficient γ_2 (fourth cumulant divided by squared variance) comes out: (i) 2, (ii) 1.2, (iii) 1.637.

Under the distribution assumption, x_1 and x_2 are independently normally distributed with means 0 and $b\sigma\sqrt{2/3}$ and with the same variance σ^2. Expressing U in terms of x_1 and x_2, we evaluate $E(U^2)$ by integration over the (x_1, x_2)-plane. The region of integration is divided into zones, as sketched in figure 2. The inner zone is a hexagon bounded by the three pairs of parallel lines,

$$z_1 = \pm C\sigma, \qquad z_2 = \pm C\sigma, \qquad z_3 = \pm C\sigma.$$

For the rejection rule, the region outside the hexagon is divided into six zones, corresponding to the six possible combinations of three values for M and two signs for z_M. For the modification rule the region outside the hexagon is divided into twelve zones, corresponding to the above six possibilities concerning z_M when y_M is the only modified observation, plus six possibilities for choosing a pair of observations to be modified, one of the residuals being positive and the other negative.

Inside the hexagon U has the constant value $b/3$. In each of the other zones U is a linear function of x_1 and x_2. For example, the zone labeled A in the rejection-rule part of figure 2 is defined by the properties:

$$M = 3, \qquad z_3 > C\sigma,$$

and in this zone we have

$$U = \frac{b}{3} - \frac{x_2}{\sigma\sqrt{6}}.$$

The zone labeled B in the modification-rule part of figure 2 is defined by the properties:

$$M = 3, \qquad z_3 > C\sigma, \qquad \text{only } y_3 \text{ is modified,}$$

and in this zone we have

$$U = \frac{b}{3} + \frac{C}{2} - \frac{x_2}{\sigma\sqrt{6}}.$$

The zone labeled C is defined by the properties:

$$y_2 \text{ and } y_3 \text{ are modified,} \quad y_3 - y_1 > C\sigma, \quad y_1 - y_2 > C\sigma,$$

and in this zone we have

$$U = \frac{b}{3} + \frac{x_1\sqrt{3} - x_2}{\sigma\sqrt{6}}.$$

In all the above it is convenient and permissible to set $\sigma = 1$.

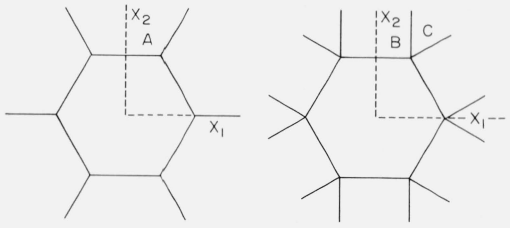

FIGURE 2. *Zones of integration.*
The diagram on the left is for the rejection rule, that on the right for the modification rule.

In principle it is possible to evaluate the double integral over each zone by expressing it as a single integral in terms of the normal integral and density functions and then using single numerical quadrature. But because that would involve much tedious detail in rotating axes, it seemed cheaper to use double numerical quadrature, integrating first for x_2 and then for x_1 by Simpson's rule. Integration was carried out over a square area of the plane, so that x_1 and x_2 ranged 5.5σ above and below their means. The integrand is well behaved within each zone but is singular on every boundary between zones. A not quite uniform grid of points was used for evaluating the integrand, so that boundaries were always encountered as end points of individual applications of Simpson's rule, never as interior points. The interval width in x_1 and x_2 of 0.10σ (or less as needed to hit the boundaries cleanly) was found satisfactory. The program was tested by finding the expectation of simple random variables whose form did not change from zone to zone. The work was done at the Yale Computer Center (IBM 7040–7094 DCS).

8. References

[1] Anscombe, F. J., Rejection of outliers, Technometrics **2**, 123–147 (1960).

[2] Anscombe, F. J. and Tukey, John W., The examination and analysis of residuals, Technometrics **5**, 141–160 (1963).

[3] Dixon, W. J., Processing data for outliers, Biometrics **9**, 74–89 (1953).

[4] Huber, Peter J., Robust estimation of a location parameter, Annals of Mathematical Statistics **35**, 73–101 (1964).

[5] Jeffreys, Harold, Theory of Probability, Oxford, Clarendon Press (1939, 1948, 1961). See sections 4.4 (all editions) and 5.7 (second and third editions, based on 5.77 in the first edition).

[6] Lieblein, Julius, Properties of certain statistics involving the closest pair in a sample of three observations, J. Res. NBS **48**, 255–268 (1952) RP2311.

[7] Sand, Francis, Investigations in residual analysis and a modification of the least-squares method for multiple regression, Ph. D. thesis, Princeton University (1963).

[8] Tukey, John W., The future of data analysis, Annals of Mathematical Statistics **33**, 1–67 (1962).

(Paper 70B2–174)