American Society for Quality

Rejection of Outliers
Author(s): F. J. Anscombe and Irwin Guttman
Source: *Technometrics*, Vol. 2, No. 2 (May, 1960), pp. 123-147
Published by: American Statistical Association and American Society for Quality
Stable URL: http://www.jstor.org/stable/1266540
Accessed: 24/03/2013 22:34

Your use of the JSTOR archive indicates your acceptance of the Terms & Conditions of Use, available at
http://www.jstor.org/page/info/about/policies/terms.jsp

JSTOR is a not-for-profit service that helps scholars, researchers, and students discover, use, and build upon a wide range of content in a trusted digital archive. We use information technology and tools to increase productivity and facilitate new forms of scholarship. For more information about JSTOR, please contact support@jstor.org.

American Statistical Association and *American Society for Quality* are collaborating with JSTOR to digitize, preserve and extend access to *Technometrics.*

http://www.jstor.org

Rejection of Outliers*

F. J. ANSCOMBE

Princeton University and University of Chicago

with collaboration by Irwin Guttman

McGill University

If one reading is a long way from the rest in a series of replicate determinations, or if in a least-squares analysis one reading is found to have a much greater residual than the others, there is temptation to reject it as spurious. Numerous criteria for the rejection of outliers have been proposed and discussed during the past 100 years. They seem always to have been regarded as something like significance tests, and attention has been focussed on rejection rates. It is suggested that rejection rules are not significance tests but insurance policies, and attention would be better focussed on error variance. A detailed study is made of the effect of routine application of rejection criteria to replicate determinations of a single value. Determinations in triplicate and quadruplicate are especially considered. Complex patterns of observations are also considered, especially factorial arrangements with high symmetry, and there is a study of the correlations between residuals. Attention is focussed mainly on rejection rules appropriate when the population variance is known, but some consideration is also given to Studentized rules.

1. SPURIOUS OBSERVATIONS

Variability or dispersion in a set of observations can arise from several different sources. Suppose that, for some reason we need not go into, it is desired to investigate the height (stature) of persons employed at a particular place. Three sources of variability in the readings are:

(i) *Inherent variability*, the variability of stature that would be observed in the population even if all measurements were perfectly accurate. This variability cannot be reduced without changing the population itself, the object of study. If we are interested in the *mean* stature of the population, we may refer to the variability as "error," since it gives rise to estimation error. But the name is misleading.

(ii) *Measurement error*, the error in using the measuring instruments. If all readings are made to the nearest centimeter, the measurement error should not exceed half a centimeter, but in fact it sometimes does. One may count as a measurement error any arithmetical mistake in reducing the original notebook entries to the form in which they are quoted as observations.

* Research carried out in part 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. Reproduction in whole or in part is permitted for any purpose of the United States Government.

This content downloaded from 202.112.90.191 on Sun, 24 Mar 2013 22:34:07 PM
All use subject to JSTOR Terms and Conditions

(iii) *Execution error*, a name we may use to include any discrepancy between what we intend to do and what is actually done, other than error in the use of measuring instruments. To include in the sample of measurements the height of some person not belonging to the population, to measure something other than height, or to select a biased sample—these would be classed as errors of execution.

No observations are absolutely trustworthy. In no field of observation can we entirely rule out the possibility that an observation is vitiated by a large measurement or execution error. If a reading is found to lie a very long way from its fellows in a series of replicate observations, there must be a suspicion that the deviation is caused by a blunder or gross error of some kind. Several possible reasons why a reading might be grossly wrong can usually be thought of without difficulty. In such cases, the reading will be checked or repeated if that is possible. If not, it may be rejected as spurious because of its big residual, even though there is no other known reason for suspecting it. In sufficiently extreme cases, no one hesitates about such rejections. The question is, where should the line be drawn? One sufficiently erroneous reading can wreck the whole of a statistical analysis, however many observations there are.

Statements to the above effect have appeared many times in the literature. Even writers who have expressed total disapproval of the rejection of outliers may be found to insert a parenthetical remark, "except for obviously incorrect readings." Reasonable argument can occur, not over the permissibility of ever rejecting outliers, but over the completeness of the rejection, how far the fact of a rejection can be forgotten—as well as over the question of where precisely to draw the line. If we could be sure that an outlier was caused by a large measurement or execution error which could not be rectified (and if we had no interest in studying such errors for their own sake), we should be justified in entirely discarding the observation and all memory of it. The act of observation would have failed; there would be nothing to report. Such an observation could justly be described as *spurious*. If, on the other hand, we could be sure that an outlier was caused, not by any large error, but by some peculiarity (nonnormality) of the inherent variability of the population under study, then it might still make good sense to discard the observation from a statistical analysis based on the method of least squares, but the observation should not be forgotten. A correct statistical summary of the observations would include a report about outliers as well as an analysis of the remaining observations (cf. Kruskal, 1960).

If a rather pronounced degree of nonnormality is expected, it may be advisable to transform the observations, or to work with medians instead of means, or some other such device.* A modified least-squares method, with weights depend-

* The suggestion about medians in this context goes back to Edgeworth (1887). It is not always satisfactory in principle to make use of any such device. If the subject of study is the output of an agricultural or industrial process, where most of the variability is inherent, the mean yield is inescapably what ought to be measured, not the median yield or mean logarithmic yield (for instance). Perhaps it should be explicitly stated that in this paper we are concerned only with observations whose primary purpose is the estimation of location parameters of some sort. Observations made primarily for the measurement of dispersion are not considered.

This content downloaded from 202.112.90.191 on Sun, 24 Mar 2013 22:34:07 PM
All use subject to JSTOR Terms and Conditions

ing on the residuals, has been suggested by Jeffreys, with particular reference to seismology and astronomy. Although such a technique is possible as a guard against spurious observations, it will not be considered here. *We envisage a situation where the intrinsic variability and ordinary measurement errors result in a nearly enough normal pattern of variation for simple averages or equal-weighted least-squares estimates to be attractive, but where it is feared that occasionally an observation will be affected by a gross error.*

An observation with an abnormally large residual will be referred to as an *outlier*. Other terms in English are "wild," "straggler", "sport" and "maverick"; one may also speak of a "discordant," "anomalous" or "aberrant" observation. *Spurious* will mean: affected by an abnormally large measurement or execution error. Usually it is not possible to check directly whether an observation is spurious; it is a matter of conjecture.

The object of this paper is to study the effect of unthinking routine use of a specified rejection rule on the estimation of means (or linear combinations of means). There are circumstances where the problem of spurious observations can appropriately be met by an impartial rejection procedure, which is thought of as part of the experimental technique. It is not suggested that the problem should always be met this way, but the results of the present study should be helpful (though not decisive) to anyone who wishes to judge individual outliers "on their merits," intelligently.

The gist of the paper can be had by reading as far as Section 3, skimming Sections 4 and 5, and then turning to Section 10. Section 7 on complex designs is to some extent a digression, and may have interest apart from the subject of outliers.

2. Sketch of the History of Rejection of Outliers

The subject of rejection rules has been held to be important and interesting in many sciences—astronomy, geodesy, chemistry, physics, ballistics. Nowadays it may be thought of as lying within the broader subject of "data processing," an essential though perhaps too little considered branch of statistical analysis. It has been discussed in countless books on the combination of observations by least squares, over a period of nearly a hundred years, as well as in many contemporary books on statistics.

One of the first references to the rejection of outliers seems to have been a remark by the leader of the German school of astronomers, Bessel, in a geodetic work published in 1838, to the effect that he had never rejected an observation merely because of its large residual; all completed observations, with equal weight, ought to be allowed to contribute to the result. "We have believed that only through strict observance of this rule could we remove arbitrariness from our results."

The first attempt at a rejection criterion based on some sort of probability reasoning was that of Peirce (1852). Peirce's argument was reproduced by Chauvenet (1863), who then gave a similar rule based on a simpler argument. A heated discussion was provoked by these rules, and by some other suggestions put forward for disposing of outliers. In due course the topic became standard

This content downloaded from 202.112.90.191 on Sun, 24 Mar 2013 22:34:07 PM
All use subject to JSTOR Terms and Conditions

in books dealing with least squares; lengthy historical surveys were given by Czuber (1891) and Wellisch (1909). See also Rider (1933).

The next significant step after Peirce and Chauvenet was apparently made by Wright (1884). After an excellent general discussion of the problems raised by the occurrence of outliers among astronomical readings, he suggests that the best rule for a computer to follow who is not the observer is to reject any observation whose residual exceeds in magnitude five times the probable error (i e. 3.37 times the standard deviation). The reason given for this is that if the Gaussian law of error is truly satisfied, only about one observation in a thousand will be rejected, "and therefore little damage will be done in any case."

From 1925 onwards statisticians have paid much attention to the subject. Student (1927) gave an interesting account of analytic chemical determinations, and proposed the use of a range criterion. Thompson (1935) may be said to have "Studentized" Wright's rule. A good survey of this literature has been given by Grubbs (1950). A recent development is the Bliss-Cochran-Tukey rule (1956), developed for use with several of the bioassays in the U. S. Pharmacopeia.

Rejection of outliers seems to have been a peculiarly American topic. At any rate, Peirce, Chauvenet and Wright were Americans, and most of the recent developments have occurred in the U. S. The subject arose in the context of least squares, or, as current jargon has it, Model I analysis of variance; and it is curious that (apparently) it was not considered by Gauss.

Chauvenet, after he had reproduced Peirce's argument, introduced his own simplified version of the rule by some remarks beginning: "The above investigation of the criterion involves some principles, derived from the theory of probabilities, which may seem obscure to those not familiar with that branch of science." The same may be said of Chauvenet and of everyone else; and familiarity with that branch of science does not remove the obscurity. It is easy now to laugh at Peirce and Chauvenet, and at other eminent nineteenth century figures such as Glaisher and Bertrand, as they wallow in the probability quagmire. But have today's statisticians been so much clearer headed? All published proposals for rejection criteria, based on any kind of mathematical reasoning, from Peirce's onwards, have an unexplained starting point or objective, presented as though it were the only obvious one and in fact utterly obscure.

All writers seem to have thought of the problem as something in the nature of significance testing.* They have varied in the significance level judged appropriate for action, but all have regarded significance levels (or something of the sort) as relevant beyond discussion. Peirce and Chauvenet thought in terms of something like a 50% significance level per set of data. If applied to a simple sample of n observations drawn from a homogeneous normal source, either criterion rejects as often as not, very roughly, so the rejection rate is

* Jeffreys puts this the other way round and suggests that the general use of tail-area probabilities by statisticians for significance testing goes back to Chauvenet's criterion (Jeffreys, 1939, p. 316; 1948, p. 357).

This content downloaded from 202.112.90.191 on Sun, 24 Mar 2013 22:34:07 PM
All use subject to JSTOR Terms and Conditions

something like 1 per $2n$ observations. Wright's rule rejects at something like the rate of 1 per 1000 observations, almost independently of n. Most modern statisticians, bemused by 5%, give rules having rejection rates of about 1 per $20n$ observations. No one has explained why this should be so. No one seems to have asked.

3. The Nature of Rejection Rules

Rejection rules are not significance tests. In a study of whether, and how often, clearly spurious observations occur in a certain field, significance tests may be appropriate. But when a chemist doing routine analyses, or a surveyor making a triangulation, makes routine use of a rejection rule, he is not studying whether spurious readings occur (he may already be convinced they do sometimes), but guarding himself from their adverse effect. The same is often true of a statistician when he performs least-squares analyses, especially if he has a single very large bulk of data or many similar smaller sets of data to analyze. The statistician may feel little temptation to exercise personal judgment in "processing" the data, and prefer to follow a rigid rule for treating outliers, whose long-term effect he knows.

A rejection rule is like a householder's fire insurance policy. Three questions to be considered in choosing a policy are

(1) What is the premium?

(2) How much protection does the policy give in the event of fire?

(3) How much danger really is there of a fire?

Item (3) corresponds to the study of whether spurious readings occur in fact—a study that is hardly possible unless plenty of readings are available. The householder, satisfied that fires *do* occur, does not bother much about (3), provided the premium seems moderate and the protection good. In what currency can we express the premium charged and the protection afforded by a rejection rule? That depends on the purpose of the observations; an answer can be given as soon as a suitable loss function is specified.

Often estimation errors are only one among several types of error affecting the whole investigation or process. For example, the error made by the manufacturer in assaying a drug is not the only reason why the patient receives an incorrect dose. The doctor judges only roughly how much of the drug the patient needs, the prescription is filled by the pharmacist with less than infinite precision, and the quantity consumed by the patient may bear only a faint resemblance to the instruction of one teaspoonful every four hours. So long as the assay errors are on the whole small, there would be no interest in very detailed information about their distribution; the bias and variance will suffice. In research experiments, the immediate object may be to measure responses to certain treatments, but the ultimate object is usually to throw light on some broad class of phenomena. Estimation errors are not the only impediment to clear perception. If the investigation is in the end abortive, estimation errors will most likely not be solely responsible. Again, we do not need to know more about the estimation errors than their variance (and bias, if any).

So variance will be considered here, although in principle any other measure

This content downloaded from 202.112.90.191 on Sun, 24 Mar 2013 22:34:07 PM
All use subject to JSTOR Terms and Conditions

of expected loss could be used. The premium payable may then be taken to be the percentage increase in the variance of estimation errors due to using the rejection rule, when in fact all the observations come from a homogeneous normal source; the protection given is the reduction in variance (or mean squared error) when spurious readings are present. Rejection rates are of no more than incidental interest.

With this approach, any suggested rejection rules can be investigated and compared. As in all previous published work on outliers, it will be assumed here that the following conditions are satisfied (they need not be in practice).

(i) Whatever circumstance causes an observation to be spurious is not expected to affect neighboring observations; all observations are supposed independent in this respect. It is also supposed that spuriousness is uncorrelated with the reading that would have been obtained had the observation been made without abnormal error.

(ii) Computation costs can be ignored. If that were not so, the "premium" would have to include the extra computation cost resulting from using the rule.

(iii) No prior knowledge concerning the means or regression coefficients that are to be estimated from the data is incorporated in the rejection rule, which is therefore "impartial." For example, suppose the observations consist of just two replicate readings, the population mean is to be estimated and the population variance is known. Any impartial rejection rule must lead to rejection if the difference between the readings exceeds some critical value, and then since there is no way of saying which is the better observation both readings must be rejected (or one after flipping a coin, perhaps). In practice, the observer may consider one of the readings reasonable and the other unlikely, and so he will retain the first and reject the other. He will then be using prior information about the population mean.

This restriction to an impartial rejection rule has an important effect on the character of our results. We return to consider it further in Section 7.

It is natural to consider rejection criteria based on the magnitude of the residual. Such a type of criterion seems likely to work better than other relatively easy criteria based on order statistics. In the least-squares analysis of complex patterns of observations, the computation of residuals is, or should be, a standard procedure, and a rejection criterion based on residuals is therefore particularly convenient. To begin with, we consider the simplest possible case, appropriate to typical chemical analyses, where only one mean is to be estimated and the population variance can be supposed known. We then pass on to complex patterns, and also consider "Studentized" criteria for use when the population variance is unknown.

Notation. Throughout, the observations will be denoted by $y_1, y_2, \cdots, y_n$, there being n observations in all. The residuals (before rejection of any outliers) will be denoted by $z_1, z_2, \cdots, z_n$. ν will always denote the number of residual degrees of freedom, i.e. ν is the rank of the matrix transforming (y_i) to (z_i).

4. Formulation of the Problem for a Simple Sample

We are given observations $y_1, y_2, \cdots, y_n$ ($n \geq 3$). It is hoped they are a

This content downloaded from 202.112.90.191 on Sun, 24 Mar 2013 22:34:07 PM
All use subject to JSTOR Terms and Conditions

random sample from a normal population $N(\mu, \sigma^2)$, where σ is known and μ is to be estimated. But possibly one or more of the y_i are spurious, coming from a different source, and ought to be rejected. We consider the effect of applying a rejection rule routinely to samples of fixed size n. Let

$$z_i = y_i - \bar{y}, \quad n\bar{y} = \sum_i y_i \qquad (i = 1, 2, \cdots, n).$$

If y_i is omitted, the average of the remaining observations is

$$\sum_{j \neq i} y_j/\nu = \bar{y} - z_i/\nu, \tag{4.1}$$

where $\nu = n - 1$. More generally, if several observations are omitted, say $y_1, y_2, \cdots, y_r$, the average of the rest is

$$\bar{y} - (z_1 + z_2 + \cdots + z_r)/(n - r). \tag{4.2}$$

Let M be the serial number of the observation having the greatest residual, so that

$$|z_M| > |z_i| \quad \text{for all} \quad i \neq M. \tag{4.3}$$

(We suppose that the observations are recorded to sufficient decimal places for no two residuals to be equal in magnitude.)

We propose to reject any observation whose residual is excessively large. The following type of rule is unsatisfactory:

RULE 0. *For given C, reject every observation y_i such that*

$$|z_i| > C\sigma.$$

Estimate μ by the mean of the retained observations.

The reason is that a single outlier, if it outlies sufficiently, can cause all the $|z|$'s to exceed $C\sigma$, and the whole sample would then be rejected. The following rule is more cautious:

RULE 1. *For given C, reject y_M if $|z_M| > C\sigma$; otherwise no rejections. Estimate μ by the mean of the retained observations, thus*

$$\begin{aligned} \hat{\mu} &= \bar{y} \quad \textit{if} \quad |z_M| < C\sigma, \\ &= \bar{y} - z_M/\nu \quad \textit{if} \quad |z_M| > C\sigma. \end{aligned}$$

Under this rule, not more than one observation can be rejected. The most frequent values for n in chemical analysis are 3 and 4, and then Rule 1 is probably the best that can be suggested. If more than one observation out of a very small sample appeared to be spurious, the observer would most likely wish to scrap them all. For large samples, however, the possibility of multiple selective rejections needs to be considered, and the following rule is suggested:

RULE 2. *Apply Rule 1. If an observation is rejected, consider the remaining observations as a sample of size $n - 1$ and apply Rule 1 again; and so on. Estimate μ by the mean of the retained observations.*

This content downloaded from 202.112.90.191 on Sun, 24 Mar 2013 22:34:07 PM
All use subject to JSTOR Terms and Conditions

It would be possible for the values of C to differ in the successive applications of Rule 1, but there is no obvious advantage in this, and, in so far as Rule 2 is considered below, C will be supposed constant. In fact, it is difficult to study Rule 2 exactly, apart from Monte Carlo computation. Rule 1 is easier, and sometimes has almost the same effect as Rule 2—namely, when there is not more than one spurious observation present in the sample, C is not very small and n is not very large.

5. Theory for Rule 1 (Simple Sample)

To study the long-run effect of application of Rule 1 to samples of size n, we consider the distribution of $\hat{\mu}$, when the y_i are interpreted as chance variables rather than as a particular realization of chance variables. The z_i and M become chance variables too.

If there are no spurious observations, the joint distribution of $(z_1, z_2, \cdots, z_n)$ is independent of the distribution of $\bar{y}$, so $\bar{y}$ and z_M are independent. Each z_i has variance $\nu\sigma^2/n$; $(n/\nu)^{\frac{1}{2}} z_i/\sigma$ has the standard normal distribution $N(0, 1)$. Let a random variable T be defined as the following function of z_M :

$$\left.\begin{aligned} T &= 0 \quad \text{if} \quad |z_M| < C\sigma, \\ &= -(n/\nu)^{\frac{1}{2}}(z_M/\sigma) \quad \text{if} \quad |z_M| > C\sigma. \end{aligned}\right\} \tag{5.1}$$

Then $\hat{\mu} = \bar{y} + \sigma T/(n\nu)^{\frac{1}{2}}$, $\bar{y}$ and T are independent, $\mathcal{E}(\hat{\mu}) = \mu$ and

$$\text{var}\,(\hat{\mu}) = \frac{\sigma^2}{n}\left(1 + \frac{\mathcal{E}(T^2)}{\nu}\right). \tag{5.2}$$

The rejection rate (proportion of observations rejected in the long run) is

$$\frac{1}{n}\,\text{ch}\,\{|z_M| > C\sigma\} = \frac{1}{n}\,\text{ch}\,\{T \neq 0\}. \tag{5.3}$$

If an observation is spurious, it is convenient to think of it as being equal to an observation from $N(\mu, \sigma^2)$ plus an extra independent error or bias. If there is one spurious observation among a sample of n, and if the bias is large enough, the observation is almost certain to be rejected under Rule 1 or Rule 2. If there are two spurious observations, and their biases are large and the difference of their biases is large, it is almost certain that both will be rejected under Rule 2.

Suppose a particular observation, y_n say, is from $N(\mu + b\sigma, \sigma^2)$, while $y_1, y_2, \cdots, y_{n-1}$ are independent from $N(\mu, \sigma^2)$. Then $\bar{y}$ is distributed $N(\mu + b\sigma/n, \sigma^2/n)$, independently of $(z_1, z_2, \cdots, z_n)$. Under Rule 1,

$$\hat{\mu} = \{\bar{y} - b\sigma/n\} + \{\sigma T/(n\nu)^{\frac{1}{2}} + b\sigma/n\}, \tag{5.4}$$

where T is defined as before by (5.1). We have

$$\mathcal{E}(\hat{\mu} - \mu)^2 = \frac{\sigma^2}{n}\left(1 + \frac{1}{\nu}\,\mathcal{E}(T + b\sqrt{\nu/n})^2\right), \tag{5.5}$$

and the rejection rate is still given by (5.3). It is to be observed that the distri-

This content downloaded from 202.112.90.191 on Sun, 24 Mar 2013 22:34:07 PM
All use subject to JSTOR Terms and Conditions

bution of z_M , and therefore that of T, depends on b and is different from what it was before. In the absence of any rejection rule, (5.5) would be replaced by

$$\mathcal{E}(\bar{y} - \mu)^2 = \frac{\sigma^2}{n}\left(1 + \frac{b^2}{n}\right). \tag{5.6}$$

In principle, (5.2) and (5.5) can be evaluated by quadrature, though if n is higher than 4 some skill in classical n-dimensional geometry is called for. The joint distribution of the (z_i) is spherical normal in a ν-dimensional flat space. If there are no spurious observations the distribution is centered at the origin; if one (or more than one) observation has an added bias the center is away from the origin. No observation is rejected provided that

$$-C\sigma < z_i < C\sigma$$

for all i, so that the point (z_i) lies inside a region R bounded by n pairs of parallel $(\nu - 1)$-flats equidistant from the origin. For $n = 3$, $\nu = 2$, R is the interior of a regular hexagon; for $n = 4$, $\nu = 3$, R is the interior of a regular octahedron. If (z_i) lies outside R, which observation is rejected, and what sign its residual has, are determined by which one of the $2n$ faces of R is intersected by the line segment joining (z_i) to the origin. To evaluate (5.2) or (5.5) an integration must be performed over the exterior of R. For (5.2) there is great symmetry, and $\mathcal{E}(T^2)$ can be expressed, after one integration along the radius vector, as $2n$ times an integral over any one face of R. The faces are regular simplexes, and the integrand is a function only of distance from the center of the simplex. Thus for $n = 3$, $\nu = 2$, each face is a line segment, and we obtain

$$\mathcal{E}(T^2) = \frac{6}{\pi}\int_{-1/\sqrt{3}}^{1/\sqrt{3}} e^{-\frac{3}{4}C^2(1+t^2)}\left\{\frac{3}{4}C^2 + \frac{1}{1+t^2}\right\}\frac{dt}{1+t^2}.$$

For $n = 4$, $\nu = 3$, each face is an equilateral triangle and the integrand is a function of distance from its center, rather more cumbersome than the above. It is far more difficult to deal with the case of one spurious observation, since the center of the distribution no longer coincides with the center of R, and much of the symmetry is lost. Only $n = 3$ has been considered. Numerical results obtained are given below in Section 10. For further progress in the accurate evaluation of (5.2) and (5.5) it will no doubt be necessary to resort to Monte Carlo techniques.

6. Approximate Formulas

Fortunately, it seems that we do not have to wait for accurate calculations to obtain results good enough for practical use. Asymptotic approximations can be found that are easy to use and ought to be tolerably reliable, as a guide to action. Just how reliable they are can only be shown convincingly by comparing them with correct values. The comparisons in Section 10 are highly encouraging, as far as they go.

For the case of no spurious observations, the proportion of observations such that the residual exceeds $C\sigma$ in magnitude, which we shall denote by α,

This content downloaded from 202.112.90.191 on Sun, 24 Mar 2013 22:34:07 PM
All use subject to JSTOR Terms and Conditions

is given by

$$\alpha = 2\Phi(-(n/\nu)^{\frac{1}{2}}C),$$

where

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

If t_α denotes the two-tailed α-point of the standard normal distribution, we have

$$t_\alpha = (n/\nu)^{\frac{1}{2}}C. \tag{6.1}$$

α would be exactly the rejection rate if we used Rule 0. If C is very large, α is very small. Moreover, if C is large, residuals exceeding $C\sigma$ in magnitude will tend to occur singly; it will be relatively unusual for two or more observations in the same sample of n to have residuals exceeding $C\sigma$ in magnitude. This is easily proved by noting that the joint distribution of any pair of residuals is normal with correlation coefficient different from ± 1. By taking C large enough, we can make the conditional chance that $|z_j| > C\sigma$, given that $|z_i| > C\sigma$, as small as we please, and hence the conditional chance that any one of $|z_j| > C\sigma$ for $j = 1, 2, \cdots, n$, with $j \neq i$, given that $|z_i| > C\sigma$, as small as we please.

It follows that, asymptotically as $C \to \infty$, we can ignore the possibility that more than one residual in the sample will exceed $C\sigma$ in magnitude, and we obtain

$$\begin{aligned} \mathcal{E}(T^2) &\sim n \int_{|t| > t_\alpha} t^2 \phi(t)\, dt \\ &= n\{2t_\alpha \phi(t_\alpha) + \alpha\}, \end{aligned} \tag{6.2}$$

and α is asymptotically the rejection rate under Rule 1.

If C is not so large as effectively to rule out the possibility that more than one of the $|z_i|$ exceeds $C\sigma$, but nevertheless we substitute (6.2) into (5.2), it may be expected that the result will approximate the effect of Rule 2 rather than Rule 1, especially if n is fairly large. We obtain this result for the simple sample if we suppose the rejected residuals to be independent and replace the divisor $n - r$ in (4.2) by $n - 1$.

Table 1 shows $2t_\alpha\phi(t_\alpha) + \alpha$ and α as functions of t_α. The entries were readily obtained from the National Bureau of Standards table of probability functions. For comparison with the results of Section 11 it may be noted that (6.2) can be expressed alternatively in terms of the tail area of a χ^2 distribution with 3 degrees of freedom.

Suppose now that there is just one spurious observation, y_n, drawn from $N(\mu + b\sigma, \sigma^2)$, and let us suppose that b is fairly large. Then it is very probable that $M = n$. Let us assume this to be so, for the moment. Then it is easy to see that $T + (\nu/n)^{\frac{1}{2}}b$ has the following distribution: outside the interval

$$I = ((\nu/n)^{\frac{1}{2}}b - (n/\nu)^{\frac{1}{2}}C, (\nu/n)^{\frac{1}{2}}b + (n/\nu)^{\frac{1}{2}}C)$$

This content downloaded from 202.112.90.191 on Sun, 24 Mar 2013 22:34:07 PM
All use subject to JSTOR Terms and Conditions

TABLE 1

Table of $2t_\alpha\phi(t_\alpha) + \alpha$, with α shown in brackets

t_α	0	2	4	6	8
2.9	.03826 (.00373)	.03630 (.00350)	.03442 (.00328)	.03263 (.00308)	.03092 (.00288)
3.0	.02929 (.00270)	.02773 (.00253)	.02625 (.00237)	.02483 (.00221)	.02348 (.00207)
3.1	.02219 (.00194)	.02096 (.00181)	.01980 (.00169)	.01869 (.00158)	.01763 (.00147)
3.2	.01663 (.00137)	.01568 (.00128)	.01478 (.00120)	.01392 (.00111)	.01311 (.00104)
3.3	.01234 (.00097)	.01161 (.00090)	.01091 (.00084)	.01026 (.00078)	.00964 (.00072)
3.4	.00905 (.00067)	.00850 (.00063)	.00798 (.00058)	.00748 (.00054)	.00701 (.00050)
3.5	.00657 (.00047)	.00616 (.00043)	.00577 (.00040)	.00540 (.00037)	.00505 (.00034)
3.6	.00472 (.00032)	.00442 (.00029)	.00413 (.00027)	.00385 (.00025)	.00360 (.00023)

the distribution is that of a standard normal variable $N(0, 1)$; inside I the chance is concentrated at the mid-point $(\nu/n)^{\frac{1}{2}}b$. We therefore have

$$\mathcal{E}(T + (\nu/n)^{\frac{1}{2}}b)^2 = 1 + \int_I (\nu b^2/n - t^2)\phi(t)\,dt.$$

For b positive and large, we can replace the upper end-point of I by ∞, obtaining

$$\mathcal{E}(T + (\nu/n)^{\frac{1}{2}}b)^2 = 1 + (\nu b^2/n - 1)\Phi(-x) - x\phi(x), \tag{6.3}$$

where

$$x = (\nu/n)^{\frac{1}{2}}b - (n/\nu)^{\frac{1}{2}}C.$$

If (6.3) is substituted into the right-hand side of (5.5), the result will be a lower bound for $\mathcal{E}(\hat{\mu} - \mu)^2$, because we have ignored the possibility that M may be different from n. We may hope, however, that the bound will be close enough to the true value to indicate how large b needs to be for the rejection rule to give good protection.

7. COMPLEX PATTERNS OF DATA

Let us turn now to complex patterns of data. We suppose the observations $y_1, y_2, \cdots, y_n$ (if none is spurious) to be drawn from independent normal distributions having common variance σ^2 and means that are given linear functions of some unknown parameters. At first we shall suppose σ^2 known.

This content downloaded from 202.112.90.191 on Sun, 24 Mar 2013 22:34:07 PM
All use subject to JSTOR Terms and Conditions

When all the unknown parameters have been estimated by least squares, the residuals $\{z_i\}$ can be calculated.

We observe first of all that in general the residuals do not all have the same chance distribution. This is illustrated by simple linear regression: y_i is drawn from $N(\mu + \beta x_i , \sigma^2)$, where the x_i are predetermined, and let us say $\Sigma_i x_i = 0$. Then

$$z_i = y_i - \bar{y} - (\Sigma_j y_j x_j / \Sigma_j x_j^2) x_i ,$$

and we find

$$\operatorname{var}(z_i) = \left(\frac{n-1}{n} - \frac{x_i^2}{\Sigma_j x_j^2}\right)\sigma^2 .$$

Thus $0 \leq \operatorname{var}(z_i) \leq (n - 1)\sigma^2/n$, both bounds being attainable for some i and some set of $\{x_i\}$. If the $\{x_i\}$ are equal-spaced, we have

$$(n - 4)\sigma^2/n < \operatorname{var}(z_i) \leq (n - 1)\sigma^2/n,$$

and if n is not very small we shall not go far wrong if we suppose all the residuals to have the same variance, say the average value over i, which is $(n - 2)\sigma^2/n$, or $\nu\sigma^2/n$. All the residuals have exactly this variance if (and only if) there are just two different x-values, with $n/2$ observations at each.

In order to keep our considerations as simple as possible, we shall restrict attention to cases where (in the absence of spurious readings) all residuals have exactly the same variance, namely $\nu\sigma^2/n$. All ordinary factorial designs, where the different levels of each factor are replicated equally often, have this property; and also Latin squares and balanced incomplete block designs. A type of design to be excluded is a factorial design where the levels of one of the factors are not equally replicated, as in an agricultural experiment on insecticides, where it would usually be advisable to replicate the "control" treatment of no insecticide more heavily than any one of the alternative insecticide treatments under test. It is to be expected that results obtained under the assumption of equal variances will be approximately correct if the residuals have only approximately equal variances, but we shall not examine this question.

We observe next that in general not every pair of residuals has the same correlation. For some designs, some of the correlation coefficients are ±1. An example is the 3 × 3 Latin square, where the residuals are equal in sets of 3, according to the letters of the orthogonal square, and so the correlation coefficients between the 36 possible pairs of the 9 residuals are 9 of them equal to 1 and 27 of them equal to $-\frac{1}{2}$. Suppose we are given that $\sigma = 1$ and the lay-out and yields are as follows, where *0*, *1*, *2* denote the three levels of a treatment:

(*0*)	$y_1 = 13.9$	(*1*)	$y_2 = 5.9$	(*2*)	$y_3 = 6.3$
(*2*)	$y_4 = 6.0$	(*0*)	$y_5 = 5.7$	(*1*)	$y_6 = 6.4$
(*1*)	$y_7 = 6.0$	(*2*)	$y_8 = 6.3$	(*0*)	$y_9 = 4.9$

This content downloaded from 202.112.90.191 on Sun, 24 Mar 2013 22:34:07 PM
All use subject to JSTOR Terms and Conditions

The residuals are

$$\begin{array}{lll} z_1 = 2.04 & z_2 = -1.22 & z_3 = -0.82 \\ z_4 = -1.22 & z_5 = -0.82 & z_6 = 2.04 \\ z_7 = -0.82 & z_8 = 2.04 & z_9 = -1.22 \end{array}$$

Since the standard deviation of residuals is $(\nu/n)^{\frac{1}{2}}\sigma = 0.47$, the values 2.04 may be judged excessively large and indicative of a spurious reading. The problem is, which reading? If we reject any *one* of y_1, y_6 and y_8 and treat that observation as missing, the estimated value to be substituted is equal to the observed value minus n/ν ($= 4.5$) times the residual (2.04), i.e. we subtract 9.2 from *one* of y_1, y_6, y_8, and proceed to estimate row, column and treatment effects in the usual way. The new residuals are

$$\begin{array}{lll} z_1 = 0 & z_2 = -0.20 & z_3 = 0.20 \\ z_4 = -0.20 & z_5 = 0.20 & z_6 = 0 \\ z_7 = 0.20 & z_8 = 0 & z_9 = -0.20 \end{array}$$

and these are satisfactorily small. But our estimates of the row, column and treatment effects will depend sharply on which one of the three questionable observations is rejected. If we insist on an impartial rejection rule, not using any prior knowledge concerning the unknown parameters to be estimated, we cannot have any preference, and to settle the deadlock we might consider rejecting all three observations. We should then scrap the whole experiment, because (in the absence of prior knowledge) too few readings are left for the unknown parameters to be uniquely estimated.

In practice the experimenter always has some prior knowledge, which may be small compared with good observations having $\sigma = 1$ but is perhaps appreciable compared with a gross error, here estimated at 9.2. If he rejects y_6 or y_8, the fitted yield will be negative, which he may find disturbing. Apart from that, he very likely thinks that rather small row, column and treatment effects are more probable than large ones, and therefore the estimates obtained after rejecting y_1 are more plausible than those obtained after rejecting either y_6 or y_8. In fact, he may guess that just the initial digit of y_1 is spurious, and y_1 should read 3.9.

One may conclude from this example that it is unwise to try to apply an impartial routine rejection rule to data such that some pairs of residuals have correlation ± 1. It may be wise to go further and exclude designs such that some pairs of residuals have large correlations, say greater than $\frac{1}{2}$ or $\frac{2}{3}$ in magnitude, because if a spurious reading occurs it is not unlikely that another reading will be rejected instead. Thus it is of interest to know, in respect of any proposed design, what correlations the residuals will have, and this is a matter not discussed in books on the design of experiments.

Let the observations $\{y_i\}$ be represented by the $n \times 1$ column vector $\mathbf{y}$, let $\boldsymbol{\theta}$ be a $(n - \nu) \times 1$ vector of nonredundant unknown parameters and let $\mathbf{A}$

This content downloaded from 202.112.90.191 on Sun, 24 Mar 2013 22:34:07 PM
All use subject to JSTOR Terms and Conditions

be the $n \times (n - \nu)$ matrix of known coefficients, such that, in the absence of spurious readings,

$$\mathbf{y} \text{ has a spherical normal distribution with } \mathcal{E}(\mathbf{y}) = \mathbf{A\theta}. \tag{7.1}$$

$\mathbf{A}$ is of rank $n - \nu$, and $\mathbf{V}$ exists such that $\mathbf{A'A} = \mathbf{V}^{-1}$. The $n \times 1$ vector of residuals $\mathbf{z}$, after least-squares estimation of $\boldsymbol{\theta}$, is given by

$$\mathbf{z} = \mathbf{Qy}, \tag{7.2}$$

where $\mathbf{Q} = \mathbf{I}_n - \mathbf{AVA'}$, $\mathbf{I}_n$ being the $n \times n$ unit matrix. $\mathbf{Q}$ has the following properties:

(i) $\mathbf{Q}$ is a $n \times n$ symmetric idempotent matrix of rank ν. It correctly transforms $\mathbf{y}$ to $\mathbf{z}$, when $\boldsymbol{\theta}$ is estimated by least squares *as if* (7.1) were true, whether or not (7.1) is true in fact. It is invariant to transformations of the parameter set $\boldsymbol{\theta}$.

(ii) The ith row of $\mathbf{Q}$ shows how a gross error in y_i is distributed among the residuals (as we see by replacing $\mathbf{y}$ by a vector consisting of 1 in the ith place and 0's elsewhere).

(iii) The variance matrix of the residuals, if (7.1) is true, is given by

$$\mathcal{E}(\mathbf{zz'}) = \mathbf{Q}\sigma^2. \tag{7.3}$$

Two practical methods of calculating $\mathbf{Q}$ are:-

(i) Solve the least-squares equations in any convenient manner and express the $\{z_i\}$ in terms of the $\{y_i\}$. For sufficiently symmetrical designs, this need be done for only one z, say z_1 .

(ii) If it is *convenient* (it is always *possible*) to choose $\boldsymbol{\theta}$ so that $\mathbf{V}$ is a scalar multiple of $\mathbf{I}_{n-\nu}$, say $\mathbf{V} = v\mathbf{I}_{n-\nu}$, then $\mathbf{Q} = \mathbf{I}_n - v\mathbf{AA'}$, and to find $\mathbf{AA'}$ we calculate scalar products of pairs of rows of $\mathbf{A}$. This method is easy for 2^k factorial designs—for which, however, it is almost the same as (i).

If all residuals have the same variance, all coefficients in the principal diagonal of $\mathbf{Q}$ are equal to ν/n, and from the relation $\mathbf{Q}^2 = \mathbf{Q}$ we easily deduce that the mean squared correlation coefficient between pairs of residuals is

$$\overline{\rho^2} = \frac{n - \nu}{(n - 1)\nu}. \tag{7.4}$$

Thus the square root of this quantity is a lower bound to the magnitude of the largest correlation coefficient between any pair of residuals. Only rarely (i.e. for a small proportion of available designs) is this lower bound attained.

If the observations have a simple two-way classification, in k rows and l columns, with $n = kl$, and the expectation of each observation y is a row constant plus a column constant, the possible correlations between residuals are easily seen to be

$$-\frac{1}{\kappa}, -\frac{1}{\lambda}, \frac{1}{\kappa\lambda}, \tag{7.5}$$

where $\kappa = k - 1$, $\lambda = l - 1$. If either k or $l = 2$, there are correlations of -1.

This content downloaded from 202.112.90.191 on Sun, 24 Mar 2013 22:34:07 PM
All use subject to JSTOR Terms and Conditions

If the observations have a one-way classification, say k observations in each of l columns, so that the expectation of each observation is just a column constant, the possible correlations between residuals are

$$-\frac{1}{\kappa}, 0; \tag{7.6}$$

again we have correlations of -1 if $k = 2$. In particular, there will be correlations of -1 if the design consists of two replications, in either one block or two blocks, of some factorial design and if all interactions between the factors are estimated, as well as the general mean and main effects and (for two blocks) the block difference.

Table 2 lists the correlations that occur among the residuals for some designs with fairly small n. Designs having correlations of ± 1 have been omitted, except for designs (2) and (14), which have been included for comparative purposes. The correlations for design (2) are an example of (7.6); those for design (3) an example of (7.5). In designs (1) and (2), three two-level factors are compared in eight observations. Both designs can be obtained by deleting four factors from a "saturated" $2^7/16$ design, as given by Plackett and Burman (1946); and it is interesting to note that two essentially different designs can be obtained in this way. Design (4) is obtained by deleting any six factors from the Plackett-Burman saturated fraction of 2^{11}. Designs (7) and (8) are half replicates of a 2^5 factorial. (7) is the obvious choice, and achieves equal-magnitude correlations. (8) illustrates the effect of another alias subgroup. The design "$2^5/2$: ABC" has correlations differing from those of (8) in sign only. Designs (9), (11) and (14) were given by Brownlee, Kelly and Loraine (1948) and have often been quoted since, as the "optimum" fractions, such that main effects are not confounded with any two-factor interactions. With respect to correlations among residuals they are not optimum, and are bettered by designs (10), (13) and (15), which, with (16) and (17), have alias subgroups permitting the lowest possible largest correlation. Design (12) is of interest, because it can be reinterpreted as two replications (in two blocks) of a 2^3 factorial, with all two-factor interactions estimated. If the three-factor interaction is also estimated, we have at once the correlation pattern of (7.5), with some correlations equal to -1. Designs (16) and (18) are essentially the same design; like (7), they achieve equal-magnitude correlations. Design (24) can be reinterpreted as a 3^3 factorial with all two-factor interactions estimated. Of all these twenty-four listed designs, (4) alone is unsymmetrical, in the sense that all the rows of $\mathbf{Q}$ do not contain exactly the same set of coefficients.

8. Formulation for Complex Patterns

Restricting attention, then, to designs such that, in the absence of spurious observations, all residuals have the same variance and no two residuals have correlation coefficient equal, or very close, to ± 1, we wish to examine the effect of routine application of a rejection rule. Two questions to consider first are: what sort of rule, and effect on what?

Let M be defined as before, by (4.3). Rules 0 and 1 of Section 4 can be adapted

This content downloaded from 202.112.90.191 on Sun, 24 Mar 2013 22:34:07 PM
All use subject to JSTOR Terms and Conditions

TABLE 2

Correlations between residuals for various designs

Design	n	ν	$\sqrt{\dfrac{n-\nu}{(n-1)\nu}}$	Correlations between residuals
(1) 2^3	8	4	0.378	$\pm\frac{1}{2}$, 0
(2) $2^3/2$ repeated: *ABC*	8	4	0.378	-1, 0
(3) 3^2	9	4	0.395	$-\frac{1}{2}$, $\frac{1}{4}$
(4) Plackett-Burman fraction of 2^5	12	6	0.302	$\pm\frac{2}{3}$, $\pm\frac{1}{3}$, 0
(5) Balanced incomplete blocks ($v = 4, k = 3$)	12	5	0.357	$-\frac{1}{2}$, $\frac{2}{5}$, $-\frac{1}{5}$, $\frac{1}{10}$
(6) 2^4	16	11	0.174	$\pm\frac{3}{11}$, $\pm\frac{1}{11}$
(7) $2^5/2$: *ABCDE*	16	10	0.200	$\pm\frac{1}{5}$
(8) $2^5/2$: *ABCD*	16	10	0.200	$\pm\frac{2}{5}$, $\pm\frac{1}{5}$, 0
(9) $2^6/4$: *ABCD, CDEF*	16	9	0.228	$\frac{5}{9}$, $-\frac{1}{3}$, $\pm\frac{1}{9}$
(10) $2^6/4$: *ABC, DEF*	16	9	0.228	$-\frac{1}{3}$, $\frac{1}{9}$
(11) $2^7/8$: *ABCD, CDEF, ACEG*	16	8	0.258	$\frac{3}{4}$, $-\frac{1}{4}$, 0
(12) $2^7/8$: *BCE, CDF, BDG*	16	8	0.258	$-\frac{3}{4}$, $\pm\frac{1}{4}$, 0
(13) $2^7/8$: *ABC, CDE, EFG*	16	8	0.258	$-\frac{1}{2}$, $\pm\frac{1}{4}$, 0
(14) $2^8/16$: *ABCD, CDEF, ACEG, EFGH*	16	7	0.293	1, $-\frac{1}{7}$
(15) $2^8/16$: *ABC, CDE, EFG, AGH*	16	7	0.293	$\pm\frac{3}{7}$, $\pm\frac{1}{7}$
(16) $2^9/32$: *ABC, CDE, EFG, AGH, BFI*	16	6	0.333	$\pm\frac{1}{3}$
(17) $2^{10}/64$: *ABE, ACF, ADG, BCH, BDI, CDJ*	16	5	0.383	$-\frac{3}{5}$, $\frac{1}{5}$
(18) Latin square	16	6	0.333	$\pm\frac{1}{3}$
(19) Balanced incomplete blocks ($v = 7, k = 3$)	21	8	0.285	$-\frac{1}{2}$, $\frac{1}{4}$, $-\frac{1}{8}$
(20) Latin square	25	12	0.212	$-\frac{1}{4}$, $\frac{1}{6}$
(21) Graeco-Latin square	25	8	0.298	$\frac{3}{8}$, $-\frac{1}{4}$
(22) 3^3	27	20	0.116	$-\frac{1}{5}$, $\frac{1}{10}$, $-\frac{1}{20}$
(23) $3^4/3$: *ABCD*	27	18	0.139	$\pm\frac{1}{6}$, 0
(24) $3^9/729$: *ABD*, *AB*^{2}*E*, *ACF*, *AC*^{2}*G* *BCH*, *BC*^{2}*I*	27	8	0.302	$-\frac{1}{2}$, $\frac{1}{4}$, $-\frac{1}{8}$

Most of the designs are factorials. Thus (10) is a quarter replicate of a factorial arrangement with 6 factors A, B, C, D, E, F, each at 2 levels; *ABC* and *DEF* are generators of the alias subgroup. (2) is a half replicate repeated in one block. (4) is given by Plackett and Burman (1946). In the balanced incomplete block designs (5) and (19), v is the number of varieties or treatment levels, k is the number of units in each block. *It is to be understood that no interactions are estimated.* Several designs can be reinterpreted as fewer-factor designs with two-factor interactions estimated, as mentioned in the text. For the incomplete block designs, interblock information is supposed not recovered.

to the present case, by changing the second sentence to read: *Estimate the unknown parameters from the retained observations by the method of least squares.* Normally it will be appropriate to perform this estimation by the two-stage method known misleadingly as "analysis of covariance," the rejected observations being regarded as missing values to be estimated.

Suppose a particular observation, y_1 say, is missing or (if present) is held

This content downloaded from 202.112.90.191 on Sun, 24 Mar 2013 22:34:07 PM
All use subject to JSTOR Terms and Conditions

to be spurious. Let $\{z_i\}$ be the residuals when some arbitrary value for y_1 is used in the least squares estimation of the parameters. Then if y_1 is replaced by $y_1 - \eta$, the residuals are replaced by $\{z_i^{(1)}\}$, where

$$z_i^{(1)} = z_i - \eta q_{i1} , \tag{8.1}$$

$\{q_{ij}\}$ being the coefficients of **Q**. The value of η which minimizes $\Sigma_i(z_i^{(1)})^2$ is

$$\hat{\eta} = \Sigma_i z_i q_{i1}/\Sigma_i q_{i1}^2 .$$

Given our assumptions, we have $\Sigma_i q_{i1}^2 = q_{11} = \nu/n$ and $\Sigma_i z_i q_{i1} = \Sigma_{ij} q_{i1} q_{ij} y_j = \Sigma_j q_{1j} y_j = z_1$. Thus

$$\hat{\eta} = (n/\nu)z_1 . \tag{8.2}$$

If all observations except y_1 are good, the correct least-squares estimates of the unknown parameters are found by substituting $y_1 - \hat{\eta}$ for y_1 in the expressions that would have been valid if all observations, including y_1 , had been good. The residuals are now given by (8.1) with (8.2) substituted, that is

$$z_1^{(1)} = 0, \qquad z_i^{(1)} = z_i - \rho_{i1} z_1 \qquad (i \neq 1),$$

where ρ_{i1} is the correlation between z_i and z_1 ; and we obtain

$$\text{var}\,(z_i^{(1)}) = (1 - \rho_{i1}^2)\nu\sigma^2/n. \tag{8.3}$$

Thus Rule 1 (which is the only rule that we consider in detail) prescribes that if $|z_M| > C\sigma$, y_M must be replaced by $y_M - nz_M/\nu$ in the parameter estimates.

How should our former Rule 2 be adapted? We have just seen that only for some rare designs are all the correlations ρ_{ij} equal in magnitude. (8.3) shows that if one observation is rejected, the new residuals will have unequal variance, in general. Provided the correlations are fairly small, it may seem reasonable—it is certainly simplest—to take no account of the changes in variance in formulating the rule, which would run: *Apply Rule 1. If an observation is rejected, compute revised residuals and apply Rule 1 again; and so on. Finally compute the least-squares estimates of the unknown parameters from the retained observations.*

Consider now how to measure the effect of routine application of one of these rules. We have argued in Section 3 above that it is appropriate to consider the sampling variances of estimates. Which estimates? In a complex design we are not necessarily interested in all the parameters that have to be estimated. The general mean and the block differences (if any) are usually of no more than secondary interest; that is, the experiment cannot be said to have been performed *in order to* estimate them. Sometimes the same is true even of the responses to some of the treatment factors. Let us therefore divide the parameters into two sets, those that are "of interest" and the remainder, and propose to study the variance matrix of the estimates of the parameters "of interest." (We shall impose a mild restriction on the choice of parameters for the "of interest" set.) Let the number of parameters "of interest" be κ. By an orthogonal transformation, the covariances of the parameter estimates can be made zero, and the product of the variances after transformation is equal

This content downloaded from 202.112.90.191 on Sun, 24 Mar 2013 22:34:07 PM
All use subject to JSTOR Terms and Conditions

to the determinant of the variance matrix before transformation. A possible measure of the "premium" charged by a rejection rule is the percentage increase in the determinant of the variance matrix caused by using the rule, when in fact no observations are spurious. We shall see that, for Rule 1 with given C, this percentage increase is proportional to κ. Rejecting an observation implies a loss of information concerning each of the parameters (in the absence of spurious observations), and in a sense it is true that the greater κ is the greater is the loss.

However, to aggregate the losses for each of the κ parameters is not altogether reasonable as a measure of the premium exacted by the rule—as we see by considering the grouping of observations. It is commonly held that to group normally distributed observations at a grouping interval not exceeding $\sigma/2$ has a negligible effect on the estimation of means, and is therefore to be recommended for ease of computation, whenever ease matters. Such a grouping increases the residual variance by about 2%, and so the sampling variance of every estimated parameter by the same percentage. If the reader considers that the recommendation about grouping is equally sound however many parameters are to be estimated, he will presumably consider the following definition of the premium charged by a rejection rule to be reasonable: *the premium is that proportional increase in σ^2 which would inflate the determinant of the variance matrix of the estimated parameters of interest by as much as the rejection rule does, when no observations are spurious.*

9. Theory for Rule 1 (Complex Patterns)

We are supposing that (in the absence of spurious readings and with no rejection rule applied)

(i) all residuals have the same variance, and

(ii) no two residuals have correlation coefficient equal or close to ±1.

We shall also suppose that

(iii) if just the κ parameters of interest are estimated, the remainder being set equal to 0, the resulting residuals all have equal variance.

These conditions concern the design—specifically, they condition the linear space spanned by the columns of $\mathbf{A}$—and they also concern the choice of the κ parameters of interest out of the total of $n - \nu$ parameters. Despite references to variance and correlations, they are independent of distribution assumptions. Condition (iii) implies, for example, that we are not allowed to pick out as the "of interest" set a one-degree-of-freedom component, e.g. the linear component, from the main effect of a three-level factor.

Let the parameters of interest be the first κ components of $\boldsymbol{\theta}$. We wish to find the ratio of determinants of the variance matrices of the estimates $(\hat{\theta}_1, \hat{\theta}_2, \cdots, \hat{\theta}_\kappa)$ calculated under the assumptions (a) that Rule 1 is applied, (b) that Rule 1 is not applied—or that Rule 1 is applied with $C = \infty$. We observe first of all that Rule 1 does not depend on the parameterization. If a nonsingular lower-triangular transformation $\mathbf{T}$ is applied to $\boldsymbol{\theta}$ and the problem is rephrased in terms of $\boldsymbol{\theta}^* = \mathbf{T}\boldsymbol{\theta}$, no change is made in the residuals or in the operation of

This content downloaded from 202.112.90.191 on Sun, 24 Mar 2013 22:34:07 PM
All use subject to JSTOR Terms and Conditions

the rejection rule, and $\mathbf{T}$ transforms the estimates of $\boldsymbol{\theta}$ to the estimates of $\boldsymbol{\theta}^*$, i.e. $\hat{\boldsymbol{\theta}}^* = \mathbf{T}\hat{\boldsymbol{\theta}}$. In particular, the leading κ-rowed minor of $\mathbf{T}$ transforms $(\hat{\theta}_1, \hat{\theta}_2, \cdots, \hat{\theta}_\kappa)$ to the corresponding estimates of interest $(\hat{\theta}_1^*, \hat{\theta}_2^*, \cdots, \hat{\theta}_\kappa^*)$. It follows that the ratio of determinants that we seek is the same for the latter estimates as for the former. There is no loss of generality in supposing that the parameters $\boldsymbol{\theta}$ satisfy the conditions

(iv) the parameters of interest are the first κ components of $\boldsymbol{\theta}$,

(v) $$\mathbf{V} = \mathbf{I}_{n-\nu}/n.$$

(If (v) is not already satisfied, apply the lower-triangular transformation $\mathbf{T}$ to $\boldsymbol{\theta}$ defined by $n\mathbf{T}'\mathbf{T} = \mathbf{V}^{-1}$.)

From (i) and (v) it easily follows that the sum of squares of the coefficients in any row of $\mathbf{A}$ is equal to $n - \nu$; (v) also implies that the columns of $\mathbf{A}$ are orthogonal and the sum of squares of coefficients in any column is equal to n. If we now introduce condition (iii), it is easy to deduce that the sum of squares of the first κ coefficients in any row of $\mathbf{A}$ is equal to κ. In the absence of any rejection rule, the least-squares estimates of $\boldsymbol{\theta}$ are

$$\hat{\boldsymbol{\theta}} = \mathbf{A}'\mathbf{y}/n.$$

Rule 1 has the effect of replacing y_M by $y_M + (n/\nu)^{\frac{1}{2}}\sigma T$, where T is defined by (5.1). Let us consider distributions conditionally on the value of M. We can apply an orthogonal transformation to the parameters of interest, so that the transformed $\mathbf{A}$ satisfies

$$a_{M2} = a_{M3} = \cdots = a_{M\kappa} = 0,$$

and then, because the transformation does not disturb the sum of squares of the first κ coefficients in any row of $\mathbf{A}$, we must have

$$a_{M1} = \sqrt{\kappa}.$$

(The sign is arbitrary; let us take the positive root.) The rejection rule does not now affect $\hat{\theta}_2, \hat{\theta}_3, \cdots, \hat{\theta}_\kappa$, but $\hat{\theta}_1$ is modified by the rejection rule by the addition of the term

$$(\kappa/n\nu)^{\frac{1}{2}}\sigma T.$$

If in fact no observations are spurious, this added term is independent of what it is added to (which has variance σ^2/n), and so we get the following result. We can transform the parameters of interest, by an orthogonal transformation depending on M, so that one new parameter has its variance changed to

$$(\sigma^2/n)\{1 + (\kappa/\nu)\mathcal{E}(T^2)\},$$

while all the other components have unaltered variances of σ^2/n. The determinant of the variance matrix (which is invariant to orthogonal transformation of parameters) is thus

$$(\sigma^2/n)^\kappa\{1 + (\kappa/\nu)\mathcal{E}(T^2)\}. \tag{9.1}$$

This result is independent of M and therefore true unconditionally. The factor

This content downloaded from 202.112.90.191 on Sun, 24 Mar 2013 22:34:07 PM
All use subject to JSTOR Terms and Conditions

in curly brackets is the ratio that we are looking for, and is invariant to any nonsingular transformation of the parameters. The premium, according to the suggested definition at the end of Section 8, is

$$\{1 + (\kappa/\nu)\mathcal{E}(T^2)\}^{1/\kappa} - 1, \tag{9.2}$$

which is less than, but may be closely equal to,

$$(1/\nu)\mathcal{E}(T^2), \tag{9.3}$$

unless $\kappa = 1$, when these expressions are exactly equal.

Now suppose that one observation, say y_n, is spurious and has bias $b\sigma$. It may happen that $M = n$, or, if not, that the first κ coefficients in the Mth row of $\mathbf{A}$ are the same as those in the nth row. Conditionally on this being so, our previous line of reasoning shows that the determinant of the matrix of mean squared estimation errors is changed from (9.1) to

$$(\sigma^2/n)^\kappa\{1 + (\kappa/\nu)\mathcal{E}(T + (\nu/n)^{\frac{1}{2}}b)^2\}. \tag{9.4}$$

This result does not hold unconditionally, but suffices for application of the results of Section 6. If no rejection rule were applied, (9.4) would be replaced by

$$(\sigma^2/n)^\kappa\{1 + (\kappa/n)b^2\}. \tag{9.5}$$

The arguments of Section 6, used to derive (6.2) and (6.3), are still valid in this context. We are now in a position to draw some conclusions.

10. Numerical Results

Both for the simple sample and for the fairly broad type of complex patterns just considered, we see, from (6.2) and (5.2) or (9.3), that the premium charged by Rule 1 can be roughly reckoned at

$$(n/\nu)\{2t_\alpha\phi(t_\alpha) + \alpha\}, \tag{10.1}$$

where $t_\alpha = (n/\nu)^{\frac{1}{2}}C$, as explained in Section 6. This should be tolerably accurate if C is so large that $n\alpha$ is quite small. If C is not so large, there is some reason to expect that (10.1) approximates the premium charged by Rule 2 rather than Rule 1.

How much premium we are willing to pay should depend on how greatly we fear spurious observations. But, as with domestic insurance, we shall probably not care very much provided the premium is small. Let us see what can be had for a premium of 2%, i.e. an effective increase of 2% in the residual variance. Setting the expression (10.1) equal to 0.02, we obtain the values of C and α given in the first part of Table 3. The lower part of the table refers to a premium of 1%. The value 1 for ν/n is not attainable, but a simple sample of large size has ν/n close to 1. The least value of ν/n for the designs given in Table 2 was 8/27, or about 0.3. Lower values for ν/n are no doubt rare, except perhaps in some screening experiments.

The protection given by Rule 1, when one observation is spurious with a large bias $b\sigma$, is given approximately by substituting (6.3) into (5.5) or (9.4) and comparing the result with (5.6) or (9.5), respectively. The best the rule

This content downloaded from 202.112.90.191 on Sun, 24 Mar 2013 22:34:07 PM
All use subject to JSTOR Terms and Conditions

could do would be always to reject the spurious observation, and in that case we should have

$$\mathcal{E}(T + (\nu/n)^{\frac{1}{2}}b)^2 = 1.$$

This is the limit of (6.3) as $b \to \infty$. To give some idea of how large b must be for the rule to work well, values of b are tabulated for which, according to (6.3),

$$\mathcal{E}(T + (\nu/n)^{\frac{1}{2}}b)^2 = 1.5.$$

The chance that the spurious observation will escape rejection is, in the cases tabulated, of the order of 0.02. For somewhat smaller b the rejection rule will be beneficial, better than no rejection rule, but the mean squared estimation error will be substantially greater than if there had been no spurious observation. If $b = (n/\nu)C$, the chance that the spurious observation will be rejected is about 0.5.

For given n and no spurious observations, as ν is reduced the effect of a rejection becomes more serious. Consequently the rejection rate α must be reduced to preserve a fixed premium, and $(n/\nu)^{\frac{1}{2}}C$ is increased. On the other hand, for a fixed bias b, $(\nu/n)^{\frac{1}{2}}b$ is reduced, and so the x in (6.3) is reduced twice over. Hence the striking increase in the b's as one reads across Table 3. If we increase the premium a little, the b's are not greatly reduced.

Thus for an experiment of fixed size n, as the design is made more ingenious and ν becomes smaller, a gross error in one of the readings spoils more estimated effects and becomes less detectable. Even if we can prevent the correlations from blowing up to ±1, it takes a bigger error to be seen.

The above findings indicate that, while the rejection rate α should depend on ν/n, it should not otherwise depend upon n. To that extent they support Wright against most other authors, who have advocated fixed rejection rates per experiment.

It remains to compare the approximate formulas with exact calculations. Com-

TABLE 3

Tabulation of rejection rules, based on approximate formulas

ν/n	1.0	0.8	0.6	0.4	0.2
2% premium					
C	3.14	2.87	2.56	2.18	1.63
α	0.00171	0.00131	0.00094	0.00058	0.00026
b	5.1	5.8	6.9	8.7	12.8
1% premium					
C	3.37	3.08	2.73	2.31	1.72
α	0.00076	0.00058	0.00042	0.00026	0.00012
b	5.4	6.1	7.2	9.1	13.3

For given premium, values of C and α are given, and also that value of b such that a bias of size $b\sigma$ in one observation inflates the mean squared estimation error only one and a half times as much as rejection of one observation at random when all observations are good.

This content downloaded from 202.112.90.191 on Sun, 24 Mar 2013 22:34:07 PM
All use subject to JSTOR Terms and Conditions

TABLE 4

Tabulation of rejection rules, based on accurate calculations

C	(n/σ^2) var $(\hat{\mu})$	Rejection rate	Approximate (n/σ^2) var $(\hat{\mu})$	Approximate rejection rate
Simple sample, $n = 3$, $\nu = 2$				
2.46003	1.04	0.002433	1.04241	0.002588
2.66184	1.02	0.001065	1.02088	0.001114
2.84623	1.01	0.000475	1.01032	0.000490
3.01724	1.005	0.000214	1.00512	0.000220
Simple sample, $n = 4$, $\nu = 3$				
2.57994	1.04		1.04134	0.002891
2.79541	1.02		1.02043	0.001247
2.99206	1.01		1.01014	0.000550
3.17434	1.005		1.00504	0.000247

The table refers to Rule 1 when there are no spurious observations, and shows values of C corresponding to given premiums of 4%, 2%, 1%, and ½%, together with the rejection rate. The last two columns show for comparison values obtained by the approximate formulas, using the C values quoted.

[Results for $n = 3$, $v = 2$, one spurious observation, are to follow.]

putations for the simple sample with $n = 3$, Rule 1, were begun by Mr. William W. Hardgrave, using a desk machine. Later, one of us (I. G.) used an IBM 650 to check and extend the calculations. Some results are presented in Table 4.

11. σ UNKNOWN

Clearly the above investigation leaves many questions unanswered. Further calculations are needed to determine the reliability of the approximate formulas. One would like to know the effect of routine application of a rejection rule to data such that the component of random variation was homogeneous but not normal—for example, having a distribution of the same shape as Student's distribution with 7 degrees of freedom (Pearson's Type VII with exponent 4), as suggested by Jeffreys, or having a "contaminated" distribution of the type considered by Tukey (1960). How does the rule function in the presence of Tukey's (1949) removable nonadditivity, or of other systematic departures from standard assumptions?

In view of statistical history during the past fifty years, the question likely to be asked first is: what happens if σ is unknown? It is not harder in principle to investigate Studentized rejection rules than those already considered. Let s^2 denote the best (quadratic unbiased) estimate of σ^2 available. If derived from the given observations (y_i) only, s^2 has ν degrees of freedom. But in general we may suppose that there is also some prior information about σ^2, equivalent to a quadratic estimate having ν_0 degrees of freedom, so that s^2 has $\nu + \nu_0$ degrees of freedom. The condition $|z_M| > C\sigma$ in Rule 1 and in the definition of T, (5.1), will be replaced by $|z_M| > Cs$.

This content downloaded from 202.112.90.191 on Sun, 24 Mar 2013 22:34:07 PM
All use subject to JSTOR Terms and Conditions

For the case of no spurious observations, we can obtain an approximate expression for the premium, corresponding to (10.1), as follows. Choosing any one of the residuals, z_1 say, for consideration, we can make a fixed orthogonal change of axes so that (z_i) is transformed to $(\zeta_1, \zeta_2, \cdots, \zeta_\nu, 0, \cdots, 0)$, where where the ζ's have independent $N(0, \sigma^2)$ distributions, and $\zeta_1 = (n/\nu)^{\frac{1}{2}} z_1$. It follows that the condition $|z_1| > Cs$ is equivalent to

$$\left(\frac{\nu(\nu + \nu_0)}{nC^2} - 1\right)\zeta_1^2 > \sigma^2\chi^2, \tag{11.1}$$

where χ^2 stands for a χ^2 variable with $\nu + \nu_0 - 1$ degrees of freedom, independent of ζ_1. The chance α that this condition is satisfied, which is approximately the rejection rate under Rule 1, is given by

$$\sqrt{\frac{nC^2}{\nu}\,\frac{\nu + \nu_0 - 1}{\nu + \nu_0 - nC^2/\nu}} = t_\alpha^{(\nu+\nu_0-1)}, \tag{11.2}$$

where the right-hand side denotes the two-tailed α-point of Student's distribution with $\nu + \nu_0 - 1$ degrees of freedom. This formula corresponds to (6.1) and extends a result given by Thompson. It can be expressed alternatively in terms of the incomplete beta function ratio tabulated by Karl Pearson, thus:

$$\alpha = I_x\left(\frac{\nu + \nu_0 - 1}{2}, \frac{1}{2}\right),$$

where $x = 1 - nC^2/\nu(\nu + \nu_0)$. The premium $\nu^{-1}\mathcal{E}(T^2)$ is given approximately by $(n/\nu\sigma^2)$ times the partial expectation of ζ_1^2, integration being confined to the region where (11.1) is satisfied. We obtain the result:

$$\frac{1}{\nu}\,\mathcal{E}(T^2) \sim \frac{n}{\nu}\, I_x\left(\frac{\nu + \nu_0 - 1}{2}, \frac{3}{2}\right), \tag{11.3}$$

where x is as above. This may be expressed alternatively in the following rule for determining C, given the premium. Multiply the premium by ν/n, and find the corresponding upper percentage point of the variance ratio (F) distribution with 3 and $\nu + \nu_0 - 1$ degrees of freedom. Calling this F, we have

$$\frac{nC^2}{\nu} \sim \frac{3F}{1 + (3F - 1)/(\nu + \nu_0)}.$$

For example, suppose that $\nu/n = 0.5$, $\nu + \nu_0 = 30$ and the premium is to be 0.02. We find that the upper 1% ($= 0.02 \times 0.5$) point of F with 3 and 29 degrees of freedom is 4.54 and so

$$2C^2 = \frac{13.6}{1.42} = 9.59, \qquad C = 2.19,$$

and from (11.2) we have $\alpha = 0.00092$. If $\nu + \nu_0 = 121$, the premium and ν/n remaining the same, we find similarly $C = 2.33$, $\alpha = 0.00079$; and for $\nu + \nu_0 = \infty$ (σ known) we have $C = 2.38$, $\alpha = 0.00076$. For a given premium, C depends quite sharply on $\nu + \nu_0$, and to a lesser extent α also varies with $\nu + \nu_0$.

For such numerical calculations, Federighi's table (1959) is useful for (11.2)

This content downloaded from 202.112.90.191 on Sun, 24 Mar 2013 22:34:07 PM
All use subject to JSTOR Terms and Conditions

and Table 16 in the Biometrika Tables (1954) for (11.3). Ingenious interpolation is called for.

It would be possible to investigate the response of the rejection rule to a spurious observation, as in Section 6, but I have not done this in general. The extreme case is the simple sample with $n = 3$, $\nu = 2$, $\nu_0 = 0$. Rule 1 is now the same as Rule 0, because at most one observation in the sample can differ from the mean by more than Cs (when $C > 1$); and it follows that formulas (11.2) and (11.3) give exact results for Rule 1. For a premium of 2% we find we must take $C = 1.154638$, and then $\alpha = 0.00667$. The rejection rule can be expressed in the form: reject either extreme observation if its distance from the median observation exceeds 82 times the distance of the other extreme observation from the median. If one of the three observations has a bias of magnitude $b\sigma$, the other two being good, the chance that the rejection rule will cause the exclusion of the bad observation is about $\frac{1}{2}$ if $b = 79$; and b needs to be three or four times this size before rejection of the bad observation can be said to be virtually certain. For all practical purposes the rejection rule is utterly useless and absurd (cf. Lieblein, 1952). One may conjecture that a Studentized rejection rule will have low power whenever $\nu + \nu_0$ is small, say less than 30.

References

Bertrand, J. (1888). *Calcul des Probabilités*. Paris. §166.

Bessel, F. W. and Baeyer, J. J. (1838). *Gradmessung in Ostpreussen*. Berlin. (*Abhandlungen von F. W. Bessel*, vol. 3, Leipzig, 1876; quoted by Czuber and Wellisch.)

Bliss, C. I., Cochran, W. G. and Tukey, J. W. (1956). A rejection criterion based upon the range. *Biometrika*, 43, 418–22.

Brownlee, K. A., Kelly, B. K. and Loraine, P. K. (1948). Fractional replication arrangements for factorial experiments with factors at two levels. *Biometrika*, 35, 268–76.

Chauvenet, W. (1863). *Manual of Spherical and Practical Astronomy*. Philadelphia. Appendix on the method of least squares, §§ 57–60. (The appendix was reprinted as *A Treatise on the Method of Least Squares*, 1868.)

Czuber, E. (1891). *Theorie der Beobachtungsfehler*. Leipzig. First part, §11.

Edgeworth, F. Y. (1887). The choice of means. *Philosophical Magazine* (5), 24, 268–71.

Federighi, E. T. (1959). Extended tables of the percentage points of Student's *t*-distribution. *Journal of the American Statistical Association*, 54, 683–8.

Glaisher, J. W. L. (1873). On the rejection of discordant observations. *Monthly Notices of the Royal Astronomical Society*, 33, 391–402.

Grubbs, F. E. (1950). Sample criteria for testing outlying observations. *Annals of Mathematical Statistics*, 21, 27–58.

Jeffreys, H. (1939). *Theory of Probability*. Oxford. §§4.4, 4.41, 5.77, 7.2. (In the second edition, 1948, the old §5.77 is recast and numbered §5.7.)

Kruskal, W. H. (1960). Some remarks on wild observations. *Technometrics*, 2, 1–3.

Lieblein, J. (1952). Properties of certain statistics involving the closest pair in a sample of three observations. *Journal of Research of the National Bureau of Standards*, 48, 255–68.

Pearson, E. S. and Hartley, H. O. (eds.) (1954). *Biometrika Tables for Statisticians*, Vol. 1. Cambridge.

Peirce, B. (1852). Criterion for the rejection of doubtful observations. *Astronomical Journal*, 2, 161–3.

Plackett, R. L. and Burman, J. P. (1946). The design of optimum multifactorial experiments. *Biometrika*, 33, 305–25.

Rider, P. R. (1933). *Criteria for Rejection of Observations* (Washington University Studies, New Series, Science and Technology, No. 8). St. Louis.

This content downloaded from 202.112.90.191 on Sun, 24 Mar 2013 22:34:07 PM
All use subject to JSTOR Terms and Conditions

"STUDENT" (1927). Errors of routine analysis. *Biometrika,* 19, 151–64.

THOMPSON, W. R. (1935). On a criterion for the rejection of observations and the distribution of the ratio of deviation to sample standard deviation. *Annals of Mathematical Statistics,* 6, 214–9.

TUKEY, J. W. (1949). One degree of freedom for nonadditivity. *Biometrics,* 5, 232–42.

TUKEY, J. W. (1960). A survey of sampling from contaminated distributions. *Contributions to Probability and Statistics: a Volume Dedicated to Harold Hotelling.* Stanford.

WELLISCH, S. (1909). *Theorie und Praxis der Ausgleichungsrechnung.* Wien. Vol. 1, §40.

WRIGHT, T. W. (1884). *A Treatise on the Adjustment of Observations by the Method of Least Squares.* New York. §§69–73.

This content downloaded from 202.112.90.191 on Sun, 24 Mar 2013 22:34:07 PM
All use subject to JSTOR Terms and Conditions