International Statistical Review, 47 (1979) 1–12 Longman Group Limited/Printed in Great Britain

Bioassay and the Practice of Statistical Inference[1]

D.J. Finney
Department of Statistics and ARC Unit of Statistics, University of Edinburgh

1 Introduction and Summary

As a branch of applied statistics, biological assay appears to have a somewhat specialized appeal. Although a few statisticians have worked in it intensively, to the majority it appears as a topic that can be neglected, either because of its difficulties or because it is trivial. Even the short course that we offer in Edinburgh as an optional component of our M.Sc. seldom attracts a student. Despite this, I am convinced that many features of bioassay are outstandingly good for concentrating the mind on important parts of biometric practice and statistical inference.

My own interest in bioassay extends over 40 years. When I joined Frank Yates at Rothamsted, among my first duties was that of assisting F. Tattersfield and his colleagues with analysing data from tests of insecticidal toxicity. A year earlier, as a student in the Galton Laboratory under R.A. Fisher, I had been introduced to probit techniques by W.L. Stevens. Confused by their complexities, I remember urging Stevens to write a book on probits; his reply was 'Why don't you write one yourself?' I also recall some years later telling Yates that I had been speculating on how the development of biometric statistics might have been changed if Fisher had chanced to be employed in pharmacology instead of agriculture. Yates disposed of my idea rapidly, and probably rightly. He held that the problems of assay would not have compared with those of agricultural experimentation in breadth of challenge or have encouraged the study of fundamental questions that became the foundation of so much of statistical inference.

In this paper, I propose to show how some aspects of bioassay illuminate the practice of statistics, especially in relation to experimental design, the use of significance tests, outliers, and other matters concerned with the interpretation of data. I first outline the logic of dilution assays and of the validity tests essential to their interpretation. The important class of radioimmunoassays provides excellent examples of problems concerning the choice of response curve, the variance of responses, the method of estimation, and the need to recognize 'second level parameters'. Experimental design can greatly affect both the adequacy and the precision of an assay, as is evident even from simple parallel line assays. Finally, I emphasize both the need for computer programs planned for routine use and the error of believing that those who are not professional statisticians require only very simple programs. Bioassay perhaps has little to offer to the more abstract theory of statistical inference, but, because some requirements are more sharply defined than in agricultural research, it can help our understanding of how to use data from experiments. Moreover, because the problems can be appreciated without much detailed explanation of the biology, bioassay is an excellent background against which to teach statistical practice.

2 What is a Biological Assay?

From the viewpoint of statisticians, bioassays are planned experiments for estimating the potencies of one or more 'stimuli' relative to a standard, using the information provided by

[1] Text slightly modified from the Eighth Fisher Memorial Lecture, delivered in Cambridge on 20 September 1978.

responses measured on biological material. Here 'stimulus' is a general term for a drug or other therapeutic preparation, a poison, or indeed any substance that can be introduced into a biological system at doses chosen by the assayist. One stimulus is designated as the *standard preparation* (S), perhaps with some degree of arbitrariness but often with the condition that it is a pure chemical compound or that supplies of it will continue to be available for a long period into the future. Other stimuli are termed *test preparations* (T). The aim is to estimate the dose of standard equal in potency to unit dose of a stated test preparation.

Underlying the estimation is the notion of a regression function of response on dose. If u is the response measured at a dose z of the standard, then

$$U_S \equiv E(u \mid z) = F(z), \tag{1}$$

where $F(\ )$ is the regression function. A particular test preparation, T, is said to have ρ times the potency of the standard if a dose z of T is equivalent to a dose ρz of S whatever the value of z; the corresponding regression function for T is therefore

$$U_T = F(\rho z). \tag{2}$$

In some circumstances, not only is this relation to be expected, but the value of ρ ought to be independent of the biological system in which responses are measured and of the definition chosen for the response. If the effect of a drug chosen as a standard can be attributed to a particular chemical entity, and if T is a preparation from another source containing the same active principle with other constituents acting simply as diluents, then ρ is a measure of relative dilution of the active principle and ought to be the same whether estimated from experiments on mice or on men or on bacterial cultures. This defines the class of *dilution assays*. Situations in which the dilution property is certain are less common than could be wished; without it, an estimate of relative potency may still be useful but its interpretation is restricted (Finney, 1947, 1978).

An assay will consist of trials of several doses of S and T, with replicates at each dose. Usually three or more doses of each preparation will be desirable; under the simplest conditions, potency estimation is possible with only two doses of S and one of T, but this can never permit validity to be tested and is seldom an acceptable procedure. All the resources of the statistician's outlook on experimental design can be used in order to optimize the precision of estimation of mean response per dose.

Equations (1) and (2) make evident that if response is plotted against log dose the two regression curves ought to be identical except for a relative shift horizontally of amount log (ρ). This is the *condition of similarity*. Many early assays, for example classical assays of vitamins, showed simple linear regression of response on log dose: for such a *parallel line assay*, the condition of similarity is that regression lines for the two preparations must be parallel. Not all response curves used for assay are so simple, but even for non-linear response curves one can speak in terms of a generalized 'parallelism' corresponding with the constant horizontal interval log ρ: the regression curves must be identical in all parameters except for that determining horizontal location.

Thus bioassay involves estimating parameters subject to a constraint of generalized parallelism, examining whether the fit is satisfactory, using the horizontal interval between response curves to estimate potency, and of course at the start using all available information in order to design the experiment optimally.

3 Validity Tests

The fundamental requirement for validity of a bioassay is the condition of similarity; the significance of any departure from it can be examined by testing the deviation from generalized parallelism. If the dose-response regression is linear and homoscedastic, a variance ratio test arises naturally from an analysis of variance. More generally, there is commonly a parameter in the regression equation that characterizes 'slope', equality of which for S and T can be examined by a general test of significance – usually with reliance upon asymptotic behaviour. One important form of response function, based upon the logistic, illustrates this well:

$$U = C+(D-C)/(1+Az^B), \tag{3}$$

a sigmoidal curve declining from D at $z = 0$ to a lower asymptote C for large z. Obviously D must be the same for both preparations, and identity of values of C is a requirement for similarity with which data will rarely conflict. The more interesting parameter B determines the slope of the curve, and a test of generalized parallelism is a test of equality of values of B. If A_S, A_T are values of the fourth parameter, the relative potency is

$$\rho = (A_T/A_S)^{1/B}. \tag{4}$$

I prefer to work with x, the logarithm of dose, and to reparameterize (3) as

$$U = C+\frac{D-C}{1+\exp\{-2(\alpha+\beta x)\}}. \tag{5}$$

The occurrence of α, β only in the expression $(\alpha+\beta x)$ shows that the transformation

$$Y = \tfrac{1}{2}\ln\{(D-U)/(U-C)\} \tag{6}$$

enables the two dose-reponse relations to be represented by parallel straight lines. Estimation must maintain the same C, D, β for S and T and use

$$\ln(\rho) = (\alpha_T-\alpha_S)/\beta, \tag{7}$$

the horizontal distance between two parallel lines.

Even when an assay is *fundamentally valid*, the validity of a particular procedure of analysis involves assumptions that also need examination. These assumptions comprise what statisticians commonly term the *model*, in my view a regrettable name; they include such statements as: 'The response curve is described by equation (5) rather than by one of the many conceivable alternatives', 'The variance is constant', 'The variance is proportional to U', and 'The distribution of individual responses is normal'. Appropriate tests of purely *statistical validity* (goodness of fit, etc.) are readily devised, though of course a test of normality (or of any other specific error distribution) is likely to have little power. The use made of the two classes of significance test, and their relation to the estimation of ρ, raise questions of general importance to statistical inference. Here is a situation quite different from that of ordinary comparative experiments, in that contrasts wanted for estimation and others wanted solely for tests of significance are logically distinct.

A fundamentally invalid assay, incompatible with equations (1) and (2), is useless – whether the explanation is that the test preparation is essentially different from the standard or that the materials have in some way been contaminated. If the test statistic were on the borderline of significance, one might not reject a single assay of a type regularly used satisfactorily, on the grounds that 1 in 20 chances do happen, but the outlook would be more severe for an isolated experiment. Tests of statistical validity pose a slightly different problem.

4 Robustness

Because of the distinction between the significance testing and the estimation aspects of bioassay, there is great scope for looking at the robustness of conclusions in relation to assumptions. Consider the commonest set of specifications and one of the simplest, that for a parallel line assay:

(i) the regression function in (1) is linear in log dose;

(ii) individual responses, u, are normally distributed about their expectations, U;

(iii) the variance of u about U is constant.

A typical assay might have 4 or 5 replicate measurements of response at each of three doses of S and T, intervals between doses being equal on a logarithmic scale. Such a symmetric design lends itself to a simple analysis of variance, with contrasts for examining parallelism and curvature as well as for estimating the parameters of

$$\left.\begin{aligned} U_S &= \alpha_S+\beta x \\ U_T &= \alpha_T+\beta x \end{aligned}\right\}, \tag{8}$$

from which equation (7) again leads to ρ.

Only a major departure from normality of the error distribution would seriously interfere with conclusions that assume normality for means. One might have greater anxiety about non-linearity or heterogeneity of variance. I have found the generalized power transformation (Box and Cox, 1964) valuable for exploring these points. The transform of the measured responses

$$y = \{(u-\delta)^i-1\}/i, \tag{9}$$

with fixed δ, i, may show a more nearly linear regression; especially for i in the range 0 to 1, the transformation also often reduces any indication of variance heterogeneity. One can easily try a number of analyses, each with a different pair of values of δ, i. Empirical trial indicates that δ has little effect unless it is taken so large as to approach the smallest recorded response. For a number of assays, I have found that, over a wide range of values of i, variance heterogeneity and linearity gave no cause for alarm and also parallelism was not challenged (Finney, 1978). For example, all three validity tests might be satisfactory for i between -0.5 and 1.2; the estimate of ρ and assessment of its precision are usually remarkably stable over a much wider range, so confirming that the quantitative conclusions are very robust in respect of uncertainties about assumptions.

Of course I am not advocating such an analysis for every assay! In an unfamiliar situation, an occasional study of sensitivity to assumptions adds to the confidence with which quantitative estimates of potency can be accepted. Where an assay technique is in regular use, a periodic check on robustness should be combined with systematic maintenance of several control charts for validity statistics.

5 Radioimmunoassay

One of the most widely used assay techniques today is that of radioimmunoassay (RIA). It is now a standard procedure for estimation of hormones and other materials in hospital patients, as an aid to diagnosis and treatment. Highly standardized experimental techniques and sophisticated autoanalysers enable hundreds of hospitals to use methods that were unknown 20 years ago, and to estimate minute quantities of important substances. A single assay may compare from 20 to 500 different test preparations with a standard, and one hospital may have two or more RIAs of a particular type per week.

The response is a count of radioactivity, arising from radioisotope labelling, that declines as dose increases; the regression of response on log dose approaches an upper limit asymptotically for small doses, and for large doses approaches a lower asymptote corresponding to a background level. Outstanding contributions to the underlying biochemical theory by Ekins and others (Ekins and Newman, 1970) indicate that the 'true' curve needs many parameters, and that perhaps expected response cannot be written as an explicit function of dose. In the routine practice of any assay, however, equation (1) should be less a mathematical description of scientific truth than an effective calibration of the observed responses. For it to fit the data well, it must come from a family with sufficient adaptability to suit all assays of a type. Equation (5) has been found very satisfactory, though to discriminate between this and an analogous equation with a different sigmoidal constituent (such as a Gaussian sigmoid) is quite impossible with any reasonable amount of data. Some categories of substances clearly show asymmetric sigmoids; I am inclined to think that introduction of an extra parameter, for example modifying equation (5) to

$$U = C + \frac{D-C}{[1+\exp\{-2(\alpha+\beta x)\}]^{\gamma}} \tag{10}$$

will attend to this complication. The constant time of counting is such as to make all counts large, almost always over 100 and frequently over 10 000. The experimental procedures inevitably introduce many sources of error other than the purely Poissonian, but once again normality does not seem to need any serious query. Now, however, account must be taken of a variance increasing markedly with the expectation of response. Rodbard [1971; Rodbard and Cooper (1970), Rodbard and Hutt, (1974)] who has probably given more thought than anyone to the details of statistical analyses for RIA, has suggested a variance function that can be written

$$\operatorname{var}(u) = V(U+HU^2); \tag{11}$$

the alternative (Finney, 1976)

$$\operatorname{var}(u) = VU^J \tag{12}$$

(where usually $0 \leqq J \leqq 2.0$) agrees at least as well with most data, is in some respects more adaptable, but can be computationally slightly slower.

Weighted least squares, the natural basis for estimation, requires iterative optimization of a standard type. I see little likelihood that empirical study of any class of assay will demonstrate that one of (11) and (12) is better for the purpose than the other, provided that suitable values of H and J can be chosen. Either function can mimic the behaviour of the other sufficiently well to give practically indistinguishable potency estimates. Indeed, with either variance function, a single assay is quite inadequate for estimating H or J at the same time as the parameters of the regression function. This fact, at first sight disturbing, may be better interpreted as meaning that the process of analysis, in respect of validity tests and potency estimation, is remarkably insensitive to variation in H or J.

6 Second Level Parameters

I have described aspects of simple parallel line assays and of radioimmunoassays in detail because I want to comment on a class of parameters. The parameters i (and δ if needed) for transformation of response (equation (9)), γ for assymetry of a curve (equation (10)), and H or J in a variance function (equation (11) or (12)) share the properties that they are needed for adequate generality yet cannot be estimated satisfactorily from the data of one experiment. Restriction to null values like $i = 1.0$, $\gamma = 1.0$, $J = 0.0$ or 1.0 would be too limiting. However, the insensitivity to the exact numerical values means that the parameters can be estimated over a series of assays of a specified type; the estimates can then be used for future assays of the same type, with confidence that robustness will protect against any dependence of conclusions upon moderate variations. In a recent paper, Miss Phillips and I described this process for J (Finney and Phillips, 1977). Our method was not theoretically flawless, and was also unnecessarily complicated for practical use, but it illustrates the possibilities.

I cannot find that anyone has explicitly recognized what I provisionally term *second level parameters*, though there is perhaps some analogy with what Barnard (1977) has recently called *labels*. By a second level parameter, I mean a quantity that is needed as a parameter for a series of experiments, but that can be taken as constant over the whole series or over a subset. Any attempt to estimate it separately in each experiment involves overparametrization. With knowledge that the results wanted are robust in relation to the exact value of the parameter, however, an estimate based on the first few experiments can thereafter be used as though it were truth. In a continuing series of routine assays, I would advocate a monitoring plan, perhaps as a control chart based upon the second level parameter, that would indicate when a sudden or gradual change in conditions made re-estimation desirable.

Some practising statisticians may find repugnant the idea of estimating a parameter from ten experiments and using this value as though it were true for the next 30 or 100 experiments. Within a well-controlled system in a hospital laboratory that undertakes good RIAs for clinical purposes, or in a pharmaceutical company that regularly assays batches of a drug, I see no difficulty in justifying the procedure. Is it in reality so different from what many of us do without conscious decision in other contexts? For example, in the analysis of cereal variety trials, the statistician will usually assume normal distribution of error with constant variance, except for the occasional experiment that declares very clearly its disagreement. I hope he does not believe in the exact truth of so palpably false a statement! Is the implicit assumption that an observation $u^{1.0}$ is normally distributed with mean $U^{1.0}$ and variance proportional to $U^{0.0}$ of any greater logical acceptability than an explicit assumption of normality for $u^{0.4}$ or a variance proportional to $U^{1.2}$, where the 0.4 or 1.2 is based on empirical evidence from like experiments? Neither is believed exactly true, and almost certainly any true description would change from experiment to experiment; both rest on substantial experience that tests of significance and estimates of other parameters are insensitive to quite large changes in values for second level parameters. For example, one RIA that I analysed using alternative values of J in equation (12) gave the results in Table 1.

Table 1

Alternative analyses of one test preparation in an oestradiol assay

J	Potency	95% limits
3.0	1.83	1.19–3.31
2.5	1.79	1.30–2.60
2.0	1.76	1.37–2.30
1.5	1.73	1.37–2.21
1.0	1.70	1.27–2.30
0.5	1.69	1.07–2.70
0.0	1.68	0.75–3.74
−0.5	1.63	0.39–6.36

Some will ask: 'If there is so much uncertainty about the parametric and distributional assumptions, why not adopt non-parametric or distribution-free methods?' This question is no more critical for bioassay than for many other fields of applied statistics. Any bioassay analysis must be parametric to the extent of consistency with equations (1) and (2), must estimate ρ and lead to probability statements about that parameter, and must provide whatever validity tests are still relevant under the lesser assumptions of the analysis. Ingenious suggestions have been made for some special circumstances. The inelegance of the calculations is not in itself a crime, but the complexity of permutations that must be examined is likely to be inordinately time consuming, especially for simultaneous assay of several test preparations or for the more sophisticated designs sometimes essential to a good experiment. I prefer the outlook that insensitivity to exact values of second level parameters justifies reasonable assumptions about them and the consequent compact and comprehensive analysis. Although there may be scope for further research, I do not see the alternative as very attractive, except to those who would have all statistical analysis distribution-free.

The concept of second level parameters is not restricted to biological assay. I believe that they, like validity tests, help to clarify inference in many branches of applied statistics. My colleague Dr H.D. Patterson is concerned with a major project of research and development on crop variety trials. These need the notion of validity of analysis and robustness in respect of assumptions about additivity and components of variance. Again, the interpretation of interactions of varietal performance with environments and seasons is made simpler by consideration of second level parameters.

7 Outliers

The problem of outliers, the occasional 'wild' values that occur in statistical data, is widespread. Much has been written about how to detect an abnormal observation and how to test the significance of its deviation from the general pattern. Such tests inevitably depend substantially upon assumptions about the error distribution. Like most statisticians, I think that in most contexts the arguments against the rules for the rejection of discrepant observations based solely on internal evidence are strong.

Assay practice raises this issue in a special form that illustrates one situation in which rejection of data can be justified. In an experiment on agricultural or industrial production, rejection of a low yield by virtue of preconceived ideas on the nature of experimental error, followed by an assertion that the treatment involved is particularly good, cannot easily be defended as an objective assessment. In bioassay, if experience and prior belief point firmly to fundamental validity, an outlier may be considered for rejection on the basis of its large deviation from an otherwise satisfactory calibration curve, or because it does not accord with other evidence of linearity or variance homogeneity, without fear that serious bias in potency estimation will ensue. The worst consequence may be an unduly optimistic statement about the precision of the estimate, and to risk this can be preferable to the alternatives of accepting a distortion of results from the anomalous response or rejecting the whole assay.

In RIA, the custom is to allot many observations (30–50) to the standard response curve in order that its position and shape shall be well-determined. For one or two responses to be very different from the general run is not uncommon, and among practitioners of RIA the view is that these ought to be rejected. Insistence on retention of all data could lead to more serious biases of estimation than a reasonable system of rejections. A little reluctantly I now accept that, in routine assays, automatic rejection of any single count that deviates from the fitted equation by, say, more than four times the estimated standard error can be preferable to a more scientific but more subjective weighing of all the evidence. However desirable the latter may be in a research study, as part of a process of clinical assessment it invites exactly the kind of unconscious subjective bias that the statistician is anxious to avoid. A practice that begins by nominating a multiple of the standard error, then rejects each count that deviates by more than this amount, and finally reanalyses the remaining data (with repetition of the cycle if necessary) is unlikely to produce systematic bias in potency estimation. The performance of such a rule must be watched: should a particular type of assay call for a large number of rejections, close examination and remedial action for improving assays would be essential.

8 Experimental Design

The treatment structure of a bioassay is often a simple factorial. Simplest of all, but limited by its inadequacy for validity testing, is that with only two dose levels of S and T, a 2^2 factorial. Important alternatives are three or four doses of S and T, to give 2×3 and 2×4 factorials, and multiple assays involving perhaps two or three or seven test preparations simultaneously that can be regarded as factorials such as 3^2, 3×4, 2×8, 3×8. Moreover, the experimental subjects (small laboratory mammals, test positions on a plate of the bacterial culture, positions in an incubator) may have a structure that calls for a block design (using litters of animals, inocula on the same plate, racks in the incubator). Except for the smallest sets of treatments, incomplete blocks will be inevitable.

If the block size is four or six or eight, there is opportunity for confounding in a manner that optimizes the useful information, remembering that a factor at four or eight levels can be formally broken down as 2^2 or 2^3. Here again, bioassay provides a distinction of logical needs not often present or recognized in other experimentation. Textbooks of design aimed at agricultural or industrial research commonly emphasize ingenious patterns that confound equally all interactions of the same order. This may be quite inappropriate in bioassay, where different interactions may carry qualitatively different information and the importance of each depends upon the knowledge available to the assayist at the start. For many types of assay that call for these complexities of design, enough is known to enable response to be measured on a scale that shows simple linear regression on log dose. The logarithm of relative potency (as in equation (7), the horizontal interval between parallel regression lines) is then estimated by a ratio of two contrasts, one the difference of mean responses or the vertical distance between the lines and the other the regression coefficient. If experience makes the assayist confident of the validity of his assay, he will be wise to confound in a way that minimizes the loss of information on these while still permitting other contrasts to give tests for parallelism and linearity that are less powerful but not of negligible utility. On the other hand, if the assayist has serious doubts about validity, he may be wise to build up experience by so confounding as to estimate validity contrasts with high precision. Imprecise potency estimates supported by powerful validity tests may be preferable to apparent precision with questionable validity. For intermediate states of knowledge, balanced confounding is not necessarily wrong but is unlikely to have special merit: there is no obvious 'exchange rate' between the values to be placed on estimates and on tests, and some treatment contrasts may lack interest from both points of view.

As a simple example, consider S and T each at three levels, in blocks of four. A balanced design would have to be unreduced and would require 15 blocks, possibly an excessive number of experimental units quite apart from whether the loss of information is wisely distributed. Table 2 shows an alternative in multiples of three blocks. This restricts confounding to the contrasts that estimate deviations from parallelism and from linearity, and the two corresponding parameters are each estimated with efficiency $\frac{3}{4}$. An assayist less confident about parallelism might prefer to sacrifice information on the linear regression coefficient in order to gain power for his validity test; a simple interchange gives Table 3, which leaves the confounding of the linearity test unaltered but transfers the other loss of information from parallelism to the estimate of regression. The rather surprising alternative in Table 4 avoids all loss of information on regression and parallelism while still losing $\frac{1}{4}$ of the information on

Table 2

Design for a (3, 3) parallel line assay in blocks of four with parallelism contrast unconfounded

	S_1	S_2	S_3	T_1	T_2	T_3
Block I	×		×	×		×
Block II	×	×			×	×
Block III		×	×	×	×	

the linearity test; this also loses information on another far less important validity test. The design in Table 5, also unsymmetric but now requiring a multiple of six blocks, loses 1/16 of the information on regression, parallelism and linearity. Evidently the designs in Tables 4 and 5 have weaknesses, but in some circumstances could be preferred to those in Tables 2 and 3 for the manner in which efficiency is distributed over contrasts.

Table 3

Design for a (3, 3) parallel line assay in blocks of four with regression contrast unconfounded

	S_1	S_2	S_3	T_1	T_2	T_3
Block	×		×	×		×
Block II	×	×		×	×	
Block III		×	×		×	×

Table 4

Design for a (3, 3) parallel line assay in blocks of four with regression and parallelism contrasts unconfounded

	S_1	S_2	S_3	T_1	T_2	T_3
Block I	×		×	×		×
Block II	×		×		× ×	
Block III		× ×		×		×

Table 5

Design for a (3, 3) parallel line assay in blocks of four with slight confounding of regression, parallelism and linearity contrasts

	S_1	S_2	S_3	T_1	T_2	T_3
Block I	×		×		× ×	
Block II		× ×		×		×
Block III	×		×	×	×	
Block IV	×		×		×	×
Block V	×	×		×		×
Block VI		×	×	×		×

This theme could be developed for larger assays, especially for those that include several test preparations. The complications are solely in initial exploration of the structure, and statistical analysis need present no difficulties. The challenge lies in the manner of putting together an assortment of blocks so as to produce a reasonable symmetry and to concentrate information where it is most needed. Bioassay can illustrate other unusual aspects of the adaptation of experimental design to satisfy constraints of the material and to meet the experimenter's requirements. For example, animal subjects may offer blocks (e.g. litters) that are not all the same size. Patterson and his co-workers (Patterson, Williams and Hunter, 1978) have recently developed designs for variety trials in unequal blocks, but otherwise this possibility has received little attention. Radioimmunoassay has yet to feel the full impact of thought on experimental design. Here experiments of 100 to 2000 'plots' are conducted with an unavoidable tempero-spatial sequence of auto-analyser processing that may produce a 'drift' or secular trend. Most practitioners are insistent that the requirements of absolute accuracy in setting up tubes and keeping records under routine conditions demand a systematic ordering that may even bring replicates to adjacent positions. The argument is strong, but analogous views in other fields have eventually been overcome in favour of controlled randomization. I cannot believe that experimental procedures using highly sophisticated apparatus are incapable of modification to incorporate this requirement, but the operational difficulties are genuine and non-negligible. Meanwhile, the best that can be done for design under the constraints deserves study.

Practitioners of RIA vary widely in apparently elementary features of design – number of doses of the standard, whether each test sample is included at one or two or more doses, and the replication at each dose. Some will have many replicates of zero dose and others scarcely any. Though familiar statistical arguments can indicate optimal allocation of available resources for precision of potency estimates, these are adequate only under ideal conditions. When precautions must be taken to detect and react to drift, outliers, poor choice of doses, and even possible errors in identification of test preparations, the position is much more complex. Any large laboratory might benefit from empirical study of the statistical performance of past assays, with a view to devising rules for effective deployment of resources in the future. An interestingly different point may arise in connection with counting time. The variance per count in a fixed counting time is an increasing function of expected count: would the overall performance of an assay be improved if the counting time were varied with dose in accordance with rules that need to be developed?

9 Quantal Response

For the purposes of this paper, little need be said about assays in which the response is *quantal*, that is to say the subject receiving a dose can be subsequently classified only as responding or not responding. The probability of response will have some form of sigmoidal relation to log dose, and this curve plays the part of the response function in equations (1), (2). The normal, the logistic, and other sigmoids have been used successfully; the only rôle of the function is to provide a smoothing of the data in the form of two parallel curves, and the exact mathematical form matters little as long as it fits adequately. Ideally, independent subjects at the same dose will show binomial variation in responses, and a general maximum likelihood analysis leads to estimation of parameters (Finney, 1978). I shall not discuss here the special problems of design.

10 Computer Programs

Surprisingly little attention has been given to producing computer programs for the analysis of biological assays. Of course any general analysis of variance program will deal with the central computations for unweighted parallel line assays, and most general least squares or maximum likelihood programs should handle RIA or quantal response assays. However, almost every type of bioassay must be seen as a routine tool for use by pharmacologists, biochemists and others, for whom it is vitally important that programs be well designed for input and output as well as for the main processing. Data must be acceptable in a format convenient for the assayist and requiring minimal change in data preparation from one assay to the next. Output must be more than just an analysis of variance and a table of means; it must complete all calculations for potency estimation and validity tests, and must present these in clear tables and diagrams.

The needs cannot be met without special programming, not intrinsically difficult but requiring careful planning. My programs PARLIN and BLISS make a start on this for the simplest unconfounded and unweighted parallel line assays and for quantal responses respectively, and have taught me how much better the job could be done. Most important is the requirement for RIA. Simplicity and speed are demanded by those who have limited appreciations of the computational complexities, yet the welfare of hospital patients depends upon the quality of the analysis. The current trend towards incorporation of microprocessors with restricted data processing facilities into the commercially available auto-analyser equipment used in processing samples and in making and recording counts has dangers. In developing a program RADIMM for personal use, I have become aware of the complexities involved in any program that does not merely apply routines uncritically to potency estimation, but instead examines validity adequately, cares for outliers, and provides output giving all relevant information for interpretation. I fear a degeneration in the standard of routine statistical analysis, if that comes to be dictated by manufacturers who ignore the need for monitoring a battery of validity tests and for assessing the precision of potency estimates. With microprocessors as cheap as they now are, it should be possible to incorporate into standard equipment a far better program; the fact that users will employ it frequently as a routine tool adds to, rather than reduces, the range of possible dangers of invalidity needing examination within the program, whether or not detailed information is output. I am currently engaged in discussions on the specification of a program; I wish I were more confident that any conclusions will be implemented.

The programming needs of bioassay and RIA should focus attention on a common misunderstanding. Too often those who are not statisticians state that methods of analysis and computer programs for their use should be short and very simple, leaving sophisticated techniques for professional statisticians. In reality, for analyses wanted frequently in routine processing of data, the opposite is nearer to the truth. The experienced statistician who scans his data carefully often senses the occurrence of non-linearities, outliers, variance heterogeneities and the like without requiring dependence on special computations and tests. A clinical biochemist is less well equipped for this type of data appreciation, and, very properly, his mind is on other features of his problem; he needs the protection of a sophisticated program that employs many components for monitoring the data as well as for the estimation, and he needs full well-formatted output. The statistician may be content with a somewhat skeletal program. A safe program for wider use is likely to be longer in terms of code and to require a larger computer for satisfactory running; it will also be more laborious to write and must be better documented. Input can be very simple, requiring minimal change from one assay to the next. A major obstacle today is that in most countries, the computers usually available for hospital laboratories are severely limited in capacity.

References

Barnard, G.A. (1977). Pivotal inference and the Bayesian controversy. *Bulletin of the International Statistical Institute.*

Box, G.E.P. and Cox, D.R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society*, **B26**, 211–252.

Ekins, R. and Newman, B. (1970). Theoretical aspects of saturation analysis. *Acta Endocrinologica*, Suppl. **147**, 11–36.

Finney, D.J. (1947). The principles of biological assay. *Journal of the Royal Statistical Society*, Suppl. **9**, 46–91.

Finney, D.J. (1976). Radioligand assay. *Biometrics*, **32**, 721–740.

Finney, D.J. (1978). *Statistical Method in Biological Assay* (3rd edition). London: Charles Griffin and Co., Ltd.

Finney, D.J. and Phillips, P. (1977). The form and estimation of a variance function, with particular reference to radioimmunoassays. *Applied Statistics*, **26**, 312–320.

Patterson, H.D., Williams, E.R. and Hunter, E.A. (1978). Block designs for variety trials. *Journal of Agricultural Science*, **90**, 395–400.

Rodbard, D. (1971). Statistical aspects of radioimmunoassays. *In* Odell and Daughaday (eds.), *Principles of Competitive Protein Binding Assays*, pp. 204–259. Philadelphia: Lippincott.

Rodbard, D. and Cooper, J.A. (1970). A model for prediction of confidence limits in radioimmunoassays and competitive protein binding assays. In *In Vitro Procedures with Radioisotopes in Medicine*, 659–674. Vienna: International Atomic Energy Agency.

Rodbard, D. and Hutt, D.M. (1974). Statistical analysis of radioimmunoassays and immunoradiometric (labelled antibody) assays: a generalized weighted, iterative, least squares method for logistic curve fitting. In *Radioimmunoassay and Related Procedures in Medicine*, vol. I, pp. 165–192. Vienna: International Atomic Energy Agency.

Résumé

Les essais biologiques sont des expériences planifiées en vue d'estimer la force d'un ou plusieurs stimuli comparativement à une norme (ou *standard*) et ce, en employant l'information fournie par les réponses d'un matériel biologique soumis à certaines mensurations (et comptages). Bien quils soient un sujet apparemment spécialisé attirant de rares statisticiens, ils sont aussi, par maints côtés remarquablement propres à faire réfléchir sur des aspects importants de la pratique de la biométrie comme de l'inférence statistique.

Dans cet article, nous montrons comment les essais biologiques peuvent jeter de la lumière sur la pratique de la statistique, en relation spécialement avec les plans d'expériences, l'emploi des tests de signification, la présence d'observations aberrantes (*outliers*), et autres sujets relatifs à l'interprétation des données. Tout d'abord nous esquissons une logique des essais de dilution, et des tests de validité nécessaires pour les interpréter. La classe importante des essais de radio-immunité fournit d'excellents exemples de problèmes se rapportant au choix de la courbe de réponse, à la variance des réponses, à la méthode d'estimation, et à la nécessité de reconnaitre 'des paramètres de second niveau'? Le plan d'expérience influe à la fois sur la pertinence et sur la précision d'un essai, ce qui peut être mis en évidence sur des exemples très simples. De nos jours, l'analyse statistique se fera ordinairement avec un calculateur. Nous insistons à la fois sur la nécessité d'un 'logiciel' qui programme pour les besoins courants et sur l'erreur qu'on commet quand on croit que des statisticiens non professionnels ont seulement besoin d'une programmation très simple.

L'essai biologique n'implique guère de théorie statistique abstraite, mais – parce que ses problèmes peuvent être jugés sans avoir de connaissances détaillées de biologie – c'est un arrière plan excellent pour l'enseignement de la statistique pratique et pour la compréhension des réalités de l'inférence statistique.