LEP strain = low egg passage strain of rabies virus
NTO = nervous tissue origin
OD = optical density
PBSA = phosphate buffered saline without Mg and Ca
PD_{50} = reciprocal of the dose which protects 50% of vaccinated animals
PM strain = Pitman-Moore strain of rabies virus
SAD strain = Street Alabama Dufferin strain of rabies virus
SRID = single radial immunodiffusion test
TCO = tissue culture origin
WHO = World Health Organization

Acknowledgments. This work was financially supported by grants of the Swiss National Foundation for Scientific Research, Impfstoffwerk Friesoythe, Burgwedel (FRG), Rhône Mérieux, Lyon (France), Schweizerisches Serum- & Impfinstitut, Bern (Switzerland), Smith-Kline R.I.T. Rixensart (Belgium) and Veterinaria AG, Zürich (Switzerland).

1 Arko, R. J., Wiktor, T. J., and Sikes, R. K., The antibody binding test for vaccine potency, in: Laboratory Techniques in Rabies, pp. 292–294. Eds M. Kaplan and H. Koprowski. WHO, Monograph Series No. 23, Geneva 1973.
2 Chappuis, G., and Tixier, G., Etude de la corrélation existant entre le test NIH et les anticorps séroneutralisants obtenus après vaccination chez le chien. Comp. Immun. Microbiol. infect. Dis. *5* (1982) 151–157.
3 Cox, J. H., Dieztschold, B., and Schneider, L. G., Rabies virus glycoprotein II. Biological and serological characterization. Infect. Immun. *16* (1977) 754–759.
4 Ferguson, M., and Schild, G. C., A single-radial-immunodiffusion technique for the assay of rabies glycoprotein antigen: application for potency tests of vaccines against rabies. J. gen. Virol. *59* (1982) 197–201.
5 Seligman, E. B., The NIH test for potency, in: Laboratory Techniques in Rabies, pp. 279–286. Eds M. Kaplan and H. Koprowski. WHO, Monograph Series No. 23, Geneva 1973.
6 Sikes, R. K., Peacock, G. V., Acha, P., Arko, R. J., and Dierks, R., Rabies vaccines: duration of immunity study in dogs. J. Am. vet. Med. Ass. *159* (1971) 1491–1499.
7 WHO Expert Committee on Rabies, pp. 25–26. Technical Report Series, No. 709, World Health Organization, Geneva 1984.
8 Pharmacopoea Helvetica ed. VII, Vaccinum rabiei ex cellulis ad usum humanum. Eidgenössische Drucksachen- und Materialzentrale, Bern 1987.
9 Pharmacopoea Helvetica ed. VII, Vaccinum rabiei inactivatum ad usum veterinarium. Eidgenössische Drucksachen- und Materialzentrale, Bern 1987.

0014-4754/88/100853-05$1.50 + 0.20/0

Robust regression in biological assay: application to the evaluation of alternative experimental techniques

A. Marazzi, C. Ruffieux and A. Randriamiharisoa

Institute of Social and Preventive Medicine, University of Lausanne, Bugnon 17, CH-1005 Lausanne (Switzerland)

Summary. Robust Huber type regression and testing of linear hypotheses are adapted to statistical analysis of parallel line and slope ratio assays. They are applied in the evaluation of results of several experiments carried out in order to compare and validate alternatives to animal experimentation based on embryo and cell cultures. Computational procedures necessary for the application of robust methods of analysis used the conversational statistical package ROBSYS. Special commands for the analysis of parallel line and slope ratio assays have been added to ROBSYS.
Key words. Biological assay; robust regression; validation of alternative assay techniques.

1. Introduction

The two major types of quantitative dose-response relations in biological assay are the *parallel line* and the *slope ratio* assay. Statistical analysis of these is based on multiple regression models and testing of linear hypotheses. Their basic assumptions and aims are summarized in Section 2, and a thorough discussion can be found in Finney[1].

Classical multiple regression analysis and linear hypotheses testing in quantitative bioassay are based on the method of Least Squares (LS, in the following) and the assumption that responses are independent random variables which are (for given doses) identically distributed according to the Gaussian (or normal) distribution. Under this assumption, the LS method is optimal (minimum variance and unbiased). However, the assumption of normality is often violated in reality. Dangerous departures from normality can be associated with the occurrence of gross errors, such as copying or keypunch errors or, more generally, occasional errors made when something indeterminate went wrong. The presence of anomalous subjects in the biological experimental material is a frequent reason for deviant responses. These departures usually show up as *outliers*, observations far removed from the majority of the data.

It is well known that the LS coefficient estimates in multiple regression are very sensitive to the presence of outliers; indeed, a single outlier can have an arbitrarily large effect on the estimates. Moreover, outliers typically inflate the standard error associated with estimates. As a consequence, the power of the classical F-test decreases drastically and the decision may completely change by removing a single distant point.

Example 1. In figure 1 a, the point P is an outlier. However, it is well aligned with the other points, and the LS

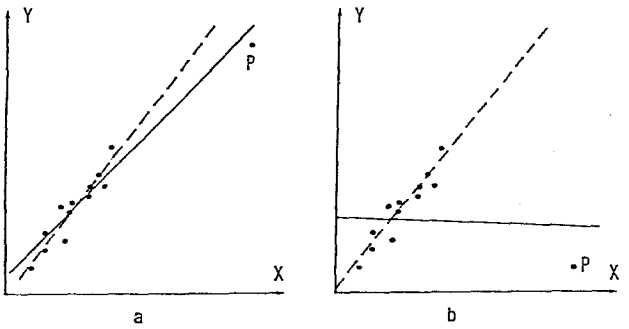

Figure 1. Effect of moving a single point on the least squares regression line.

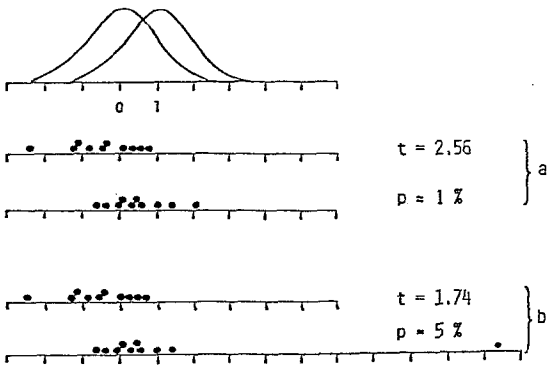

Figure 2. Effect of moving a single observation on the t-test.

regression line computed on the complete data set (continuous line) is close to the LS line which ignores the point P (broken line). In contrast, in figure 1 b, the result is strongly biased by the presence or absence of point P, and the continuous line is clearly a totally misleading summary of the relation between y and x.

Example 2. The two samples in figure 2a, are pseudo-random numbers generated according to two Gaussian distributions with the same scale parameter $\sigma = 1$ but with means of 0 and 1. A t-test detects the difference at the level 1%. However, if the highest value of the second sample is moved beyond 10.5 the test becomes non-significant at the 5% level. Paradoxically, the point is moved in a direction which should stress the difference!

Finally, it must be noted that several percent (up to 10) of outliers are rather common. An extensive discussion on the nature and frequency of gross errors, as well as detailed examples of their effects, can be found in Hampel et al.[2].

During the past two decades, considerable theoretical effort has been put into developing statistical procedures that are resistant to outliers, and stable with respect to deviations from a given (usually – but not necessarily – Gaussian) distributional model. These statistical procedures are called *robust* techniques. In particular, methods for robust regression, estimation and testing of linear models (analysis of variance) have received a lot of attention. Among them, procedures based on *M-estimators* seem to be the most successful. Their theory is developed in Hampel et al.[2] and Huber[5]. In Section 3, some elements of the theory of robust regression based on *Huber's type* M-estimators are reviewed and adapted to the problems of parallel line and slope ratio assay.

Methods and procedures used were made available in the statistical package ROBSYS[8–10]. Section 4 gives a brief account of the main features of ROBSYS.

ROBSYS has been applied to the analysis of a series of experiments carried out within the project 'Alternatives to Animal Experimentation' supported by the Swiss National Science Foundation. Their aim was to validate a number of alternative biological systems (based on embryo or cell cultures) in the framework of quantitative bio-assay. To this end, a special experimental design which ensured the fundamental *condition of similarity* of an analytic dilution assay was devised. The basic idea is to use two known concentrations of the same substance as standard and test preparation of the assay. In this way, questions of *bias* and *precision* in the estimation of the *potency* (=ratio between the two known concentrations) can be easily investigated. The experimental scheme is briefly described in Section 5. The results and the statistical analyses are discussed in Sections 6 and 7.

2. The statistical models

2.1 Assumptions

Suppose a subject receives *dose* z of a particular stimulus (substance, preparation), and that the *response* subsequently measured is a random variable U. A response which is to be of value for assay purposes must depend in some manner upon the dose. Let $E(U|z) = F(z)$, where $E(U|z)$ denotes the expected response upon z and F some real function.

Suppose that for a *standard* (S) and a *test* (T) *preparation* $E(U_S|z) = F_S(z)$ and $E(U_T|z) = F_T(z)$. (In the following, quantities corresponding to the standard preparation are indexed by the suffix S, for ex. U_S, F_S, ..., and quantities corresponding to the test preparation by the suffix T, for ex. U_T, F_T, ...). In many assays, the test preparation T, behaves as if it were simply a dilution (or a concentration) of the standard preparation S, in a diluent that is inert in respect of the response used. Obviously, for such an assay (called *analytic dilution assay*) there is a constant ϱ such that:

$$F_T(z) = F_S(\varrho z) \quad \text{for all } z \text{ (\textit{condition of similarity}).} \quad (2.1)$$

ϱ is called the *potency of T with respect to S.*

Usually, several subjects are *allocated at random* to various doses of T and S; then responses are observed. For an analytic dilution assay, the statistical problem is to estimate (the parameters of) the regression functions F_S and F_T, test the condition of similarity, check the model assumptions, and derive an estimate $\hat{\varrho}$ of ϱ. Sometimes, the statistical analysis must take into account the group-

ing of subjects into *blocks* (for example, animal litters or day of experimentation).

In this paper we consider the common models based on the following assumptions:

1. Errors of measurement of the doses are negligible.
2. For any fixed z, values of the response U are random observations distributed according to some density function $f(u|z)$ in which z plays the part of a parameter.
3. For a suitable pair of transformations $x = g(z)$ and $Y = H(U)$ uniquely defining x and Y as functions of z and U respectively, the regression of Y on x is linear (at least over a restricted dose range):

$$E(Y|x) = \alpha + \beta x. \tag{2.2}$$

x and Y are known as the *dose* and the *response metameters*.

4. For any fixed x, the response metameter Y follows a normal distribution with expected value $E(Y|x)$ and variance $V(Y|x)$.
5. The *scale parameter* of the distribution of Y is independent of x. In other words, $V(Y|x) = \sigma^2$, where σ is a constant.

In practice, the transformation g can be taken as $x = z^\lambda (\lambda \neq 0)$ or $x = \log(z)$ $(\lambda = 0)$. In these cases, it follows from the condition of similarity that:

$$E(Y_S|x) = \alpha + \beta x \quad \text{and} \quad E(Y_T|x) = \alpha + \beta \log \varrho + \beta x \quad \text{if } \lambda = 0, \tag{2.3}$$

$$E(Y_S|x) = \alpha + \beta x \quad \text{and} \quad E(Y_T|x) = \alpha + \beta \varrho^\lambda x \quad \text{if } \lambda \neq 0. \tag{2.4}$$

If $\lambda = 0$, the S and T lines are parallel (*parallel line assay*); writing $\alpha_S = \alpha$, $\alpha_T = \alpha + \beta \log \varrho$ for the intercepts and β for the common slope, one gets $\log \varrho = (\alpha_T - \alpha_S)/\beta$. If $\lambda \neq 0$, the S and T lines intersect at $x = 0$ (*slope ratio assay*); writing β_S, β_T for their slopes, one gets $\varrho = (\beta_T/\beta_S)^{1/\lambda}$.

Remark 1. In order to use classical least squares inference and analysis of variance, the response metameter Y should follow a normal distribution exactly. In other words the density of Y for given x is $f(y|x) = \phi((y - \alpha - \beta x)/\sigma)$ where

$$\phi(y) = \frac{1}{\sqrt{2\pi}} \exp[-y^2/2]. \tag{2.5}$$

Departures from this model (as discussed in Hampel et al.[2], chapter 1) are harmless if robust statistical inference is used. A fundamental kind of departure is given by the 'gross-error-model'. Instead of strictly accepting (2.5), the researcher assumes that a (known) fraction ε $(0 \leqq \varepsilon < 1)$ of the data may consist of gross errors with arbitrary (unknown) distribution $\zeta(y|x)$. The distribution underlying the observations is thus

$$f(y|x) = (1 - \varepsilon)\phi((y - \alpha - \beta x)/\sigma) + \varepsilon \zeta(y|x). \tag{2.6}$$

Remark 2. Transformation to linearity is not in itself interesting, but simplifies the analysis. On the other hand, any transformation of response alters the variance and it may be hard to simultaneously satisfy assumptions 3, 4 and 5. Naturally, one could, if necessary, instead work with the untransformed regression function and the appropriate, but more complex, fitting technique. The robust method can be correspondingly extended.

2.2 The multiple regression models

We review the basic models in the form implemented in ROBSYS.

For the *parallel line assay* we consider the *symmetric* and *unsymmetric randomized blocked* and *unblocked designs*. In other words, we suppose that a full linear model for the response metameter Y can be constructed with the help of dummy variables as follows:

$$\begin{aligned} Y = {} & \mu V + \pi W + \beta_2 B_2 + \beta_3 B_3 + \ldots + \beta_b B_b \\ & + \eta_1 X_S + \eta_2 X_S^2 + \ldots + \eta_{h-1} X_S^{h-1} \\ & + \theta_1 X_T + \theta_2 X_T^2 + \ldots + \theta_{k-1} X_T^{k-1} + \textit{error} \end{aligned} \tag{2.7}$$

where

$V \equiv 1$

$W = -1$, if Y is a response to the standard S
$\quad +1$, if Y is a response to the test T

$X_S =$ dose value for Y, if Y is a response to S
$\quad 0$, otherwise

$X_T =$ dose value for Y, if Y is a response to T
$\quad 0$, otherwise

$B_l = 1$, if Y is observed in block l
$\quad 0$, otherwise.

Here, $l = 1, \ldots, b$ and b is the number of blocks; for a design without blocks the terms $\beta_2 B_2, \beta_3 B_3, \ldots, \beta_b B_b$ are absent. The number of terms which can be included into the full model depends on the numbers of doses h and k used for S and T.

The model (2.7) can be re-expressed as follows. Suppose, without loss of generality, that $k \leqq h$ and define:

$$L_1 = X_S + X_T, L_2 = X_S^2 + X_T^2, \ldots, L_{k-1} = X_S^{k-1} + X_T^{k-1}$$

$$L_1' = X_S - X_T, L_2' = X_S^2 - X_T^2, \ldots, L_{k-1}' = X_S^{k-1} - X_T^{k-1}.$$

Then, (2.7) is equivalent to:

$$\begin{aligned} Y = {} & \mu V + \pi W + \beta_2 B_2 + \beta_3 B_3 + \ldots + \beta_b B_b \\ & + \gamma_1 L_1 + \ldots + \gamma_{k-1} L_{k-1} \\ & + \eta_k X_S^k + \ldots + \eta_{h-1} X_S^{h-1} \\ & + \delta_1 L_1' + \ldots + \delta_{k-1} L_{k-1}' + \textit{error} \end{aligned} \tag{2.8}$$

with $\gamma_j = \frac{1}{2}(\eta_j + \theta_j)$ and $\delta_j = \frac{1}{2}(\eta_j - \theta_j)$, $j = 1, \ldots, k-1$.

The model (2.8) is more convenient for testing purposes because the hypotheses of interest directly fit the canonical from (3.5) which sets some of the coefficients to 0. In practice, the model (2.8) is initially assumed. A preliminary investigation will be presumed to have established that, over a certain range, the regression functions are

(almost) linear. However, in a routine analysis, the hypothesis $H_L: \gamma_2 = \ldots = \gamma_{k-1} = \delta_2 = \ldots = \delta_{k-1} = \eta_k = \ldots = \eta_{h-1} = 0$ should be tested as a check that nothing has seriously disturbed linearity (*statistical invalidity*). Assume now that linearity has been accepted. In a well planned assay, a large difference between treatment means should not arise. (Such a difference could indicate that the responses to either the lowest or the highest doses of T lie far outside the range of responses to S. Moreover, a large difference in mean responses will decrease the precision of potency estimation – see Finney [1], Section 4.14). In order to compare treatment means we test $H_T: \pi = 0$. Deviations from parallelism can be analysed by testing $H_P: \delta_1 = 0$. If linearity and no difference between preparations have been accepted, a rejection of parallelism would indicate that the condition of similarity is violated (*fundamental invalidity*). The whole assay is then suspect, and should be discarded. Finally, in a good assay the hypothesis $H_R: \gamma_1 = 0$ should be strongly rejected.

The most common designs of *slope ratio assay* are similar to those of parallel line assay. Yet, a certain number of observations at zero dose (*blanks*) are usually included. Suppose the full regression model is

$$\begin{aligned} Y = {} & \mu V + \pi W + \alpha_0 X_0 \\ & + \eta_1 X_S + \eta_2 X_S^2 + \ldots + \eta_{h-1} X_S^{h-1} \\ & + \theta_1 X_T + \theta_2 X_T^2 + \ldots + \theta_{k-1} X_T^{k-1} \\ & + \beta_2 B_2 + \beta_3 B_3 + \ldots + \beta_b B_b + error \end{aligned} \tag{2.9}$$

where $V, X_S, X_T, B_2, \ldots, B_b$ are defined as in (2.7) and

$W = -1$, if Y is a response to the S
$\quad\quad 1$, if Y is a response to the test T
$\quad\quad 0$, if Y is a blank response

$X_0 = \quad 1$, if Y is a blank response
$\quad\quad 0$, otherwise.

b is the number of blocks; for a design without blocks the terms $\beta_2 B_2, \beta_3 B_3, \ldots, \beta_b B_b$ are absent. h and k are the number of doses used for S and T. If no blanks are available, the term $\alpha_0 X_0$ is absent.

As for the parallel line assay, linearity is a requirement for *statistical validity* of the analysis. Therefore, we usually test the hypothesis $H_L: \eta_2 = \ldots = \eta_{h-1} = \theta_2 = \ldots \theta_{k-1} = 0$ as a first step. Intersection of the regression lines at $x = 0$ is a requirement for *fundamental validity*, deriving from the hypothesis of similarity and analogous to parallelism in parallel line assay. Therefore, if linearity is accepted, we test $H_T: \pi = 0$ as a second step. In order to test whether the equation remains valid down to zero dose, one can test $H_B: \alpha_0 = 0$. In fact, a slight curvature at very low doses is not uncommon. In this case, estimation of potency from the non-zero dose levels (i.e. omitting the blanks) seems desirable. Finally, the hypothesis $H_R: \eta_1 = \theta_1 = 0$ should be strongly rejected in a meaningful assay.

3. Robust regression and testing of linear hypotheses

The reader is referred to Hampel et al.[2] and Huber [5] for a thorough discussion of this topic. This section summarizes the basic procedures we use in ROBSYS for the analysis of parallel line and slope ratio assays. Another approach to robust quantitative bioassay is developed in Sen [13, 14].

3.1 Notations

Assume that p unknown parameters $\vartheta_1, \ldots, \vartheta_p$ are to be estimated from n observations (responses) $y_1, \ldots, y_n$ to which they are related according to the model

$$\Omega: \; y_i = \sum_{j=1}^{p} x_{ij} \vartheta_j + error_i, \tag{3.1}$$

where the x_{ij} are known coefficients and the errors are independent with (approximately) identical distributions which depend on an unknown scale parameter σ. The matrix of the coefficients x_{ij} (called *design matrix*) will be denoted by $\mathbf{X}$.

Example 1. In the model (2.8), $p = h + k + b - 1$ (or $p = h + k$ for a design without blocks); the parameters and the explanatory variables (i-th value) of (2.8) correspond to the parameters ϑ_j and the coefficients x_{ij} in (3.1) according to the following scheme:

μ	π	γ_1	$\ldots$	γ_{k-1}	η_k	$\ldots$	η_{h-1}	δ_1	$\ldots$	δ_{k-1}	β_2	$\ldots$	β_b
V	W	L_1	$\ldots$	L_{k-1}	X_S^k	$\ldots$	X_S^{h-1}	L_1'	$\ldots$	L_{k-1}'	B_2	$\ldots$	B_b
ϑ_1	ϑ_2	ϑ_3	$\ldots$	ϑ_{k+1}	ϑ_{k+2}	$\ldots$	ϑ_{h+1}	ϑ_{h+2}	$\ldots$	ϑ_{h+k}	ϑ_{h+k+1}	$\ldots$	$\vartheta_{h+k+b-1}$
x_{i1}	x_{i2}	x_{i3}	$\ldots$	$x_{i,k+1}$	$x_{i,k+2}$	$\ldots$	$x_{i,h+1}$	$x_{i,h+2}$	$\ldots$	$x_{i,h+k}$	$x_{i,h+k+1}$	$\ldots$	$x_{i,h+k+b-1}$

Example 2. In the model (2.9) $p = h + k + b$ (or $p = h + k + 1$ for a design without blocks); the parameters and the explanatory variables (i-th value) of (2.9) correspond to the parameter ϑ_j and the coefficients x_{ij} in (3.1) according to the following scheme:

μ	π	α_0	η_1	$\ldots$	η_{h-1}	θ_1	$\ldots$	θ_{k-1}	β_2	$\ldots$	β_b
V	W	X_0	X_S	$\ldots$	X_S^{h-1}	X_T	$\ldots$	X_T^{k-1}	B_2	$\ldots$	B_b
ϑ_1	ϑ_2	ϑ_3	ϑ_4	$\ldots$	ϑ_{h+2}	ϑ_{h+3}	$\ldots$	ϑ_{h+k+1}	ϑ_{h+k+2}	$\ldots$	ϑ_{h+k+b}
x_{i1}	x_{i2}	x_{i3}	x_{i4}	$\ldots$	$x_{i,h+2}$	$x_{i,h+3}$	$\ldots$	$x_{i,h+k+1}$	$x_{i,h+k+2}$	$\ldots$	$x_{i,h+k+b}$

3.2 Parameter estimation

Classically, the problem of estimating the parameters $\vartheta_1, \ldots, \vartheta_p$ is solved by minimizing the sum of squares $\sum_i (y_i - \sum_j x_{ij}\vartheta_j)^2$ or, equivalently, by solving the system of p 'normal equations' obtained by differentiating this sum with respect to the parameters and equating the partial derivatives to 0. The scale parameter σ is then estimated by $\hat{\sigma}_{LS} = \left(\frac{1}{n-p}\sum r_i^2\right)^{1/2}$, where the r_i are the residuals with respect to the LS fit.

Various procedures have been proposed to obtain robust parameter estimates. We follow a proposal due to P. J. Huber. First, suppose that σ is known. In order to bound the influence of gross-errors on the parameter estimates we minimize

$$Q(\vartheta_1, \ldots, \vartheta_p) = \sum_i \varrho((y_i - \sum_j x_{ij}\vartheta_j)/\sigma), \tag{3.2}$$

where ϱ is a convex function of the residuals, which increases less rapidly than the squared function. By taking partial derivatives of θ with respect to $\vartheta_1, \ldots, \vartheta_p$, and defining $\psi = \varrho'$, we obtain the following system of equations

$$\sum_i^n \psi((y_i - \sum_j x_{ij}\vartheta_j)/\sigma)x_{ik} = 0, \quad k = 1, \ldots, p. \tag{3.3}$$

for $\vartheta_1, \ldots, \vartheta_p$. The solution $\hat{\vartheta}_1, \ldots, \hat{\vartheta}_p$ of (3.3) is called a *Huber estimate* of $\vartheta_1, \ldots, \vartheta_p$.

The initial, the function ϱ is arbitrary and must be chosen. On theoretical grounds, one of the best choices is:

$$\varrho(r) = \begin{cases} c|r| - c^2/2 & \text{for } |r| \geqq c \\ r^2/2 & \text{for } |r| < c \end{cases}$$

which corresponds to $\psi(r) = \max(-c, \min(c, r))$. This choice is the solution of various optimality problems, which can be compared to the following insurance problem. The LS procedure is optimal (that is, it has the highest 'precision' or, equivalently, the lowest variance) for the Gaussian model. It happens to be nonrobust and we want to insure against accidents caused by deviations from the model. We clearly have to pay for it by sacrificing some precision at the Gaussian model. The questions are, of course, how much precision we are willing to sacrifice, and against how large a deviation we would like to insure. The parameter c is related to this price and to the fraction of outliers in (2.6). For example, for $c = 1.35$ (the most common value used in practice) the Huber estimate is about 95% as precise as the LS estimate for Gaussian errors. However, it is the most precise (maximum likelihood) estimate with respect to the worst distribution (i.e. the "least favourable" distribution in a two persons game between the Statistician and Nature) of the form (2.6) with $\varepsilon \approx 5\%$ and performs well (i.e. robustly and precisely) for all distributions of this class. In the limit case $c \to \infty$, the Huber estimate tends to the LS estimate, which is optimal for $\varepsilon = 0\%$ and for which no insurance is paid.

In practice, σ is unknown. In order to estimate it, a supplementary equation for σ

$$\sum_i \chi((y_i - \sum_j \vartheta_j x_{ij})/\sigma) = (n-p)\,BETA, \tag{3.4}$$

is solved simultaneously with (3.3). The solution $\hat{\sigma}$ is a robust estimate of σ. Here, $\chi(r) = \psi(r)^2/2$ and the constant *BETA* is set so that $\hat{\sigma}$ is asymptotically (i.e. for n large) unbiased for Gaussian errors. (More precisely, $BETA = \int \chi(r)\,\phi(r)\,dr$.) Notice again that $\hat{\sigma}$ reduces to $\hat{\sigma}_{LS}$ for $c \to \infty$.

Finally, the covariance matrix of the parameter estimates $\hat{\vartheta}_1, \ldots, \hat{\vartheta}_p$ is estimated by $\hat{\sigma}^2 f(\mathbf{X}^T\mathbf{X})^{-1}$, where f is an appropriate correction factor. (More precisely, $f = [\int \psi^2(r)\,\phi(r)\,dr]/[\int \psi'(r)\,\phi(r)\,dr]^2$.)

Remark. The solution of (3.3) and (3.4) cannot be symbolically derived but can be computed numerically. The better known algorithms are based on the Newton method, and several of them are implemented in ROBSYS. We give an illustration in figure 3. Diagram 'a' shows some data points and the least squares regression line LS_0. This line is clearly unsatisfactory; nevertheless, we use it as the starting value for an iterative process. In diagram 'b', a band centered on LS_0 has been indicated by two parallel lines. The band width is not essential in this example, but it usually depends on the data dispersion. In the same diagram, the points outside the band have been translated, parallel to the y axis, to the border of the band. In this way, new pseudo-observations can be defined: they coincide with the original points inside the band and with the 'modified points' on the border. Dia-

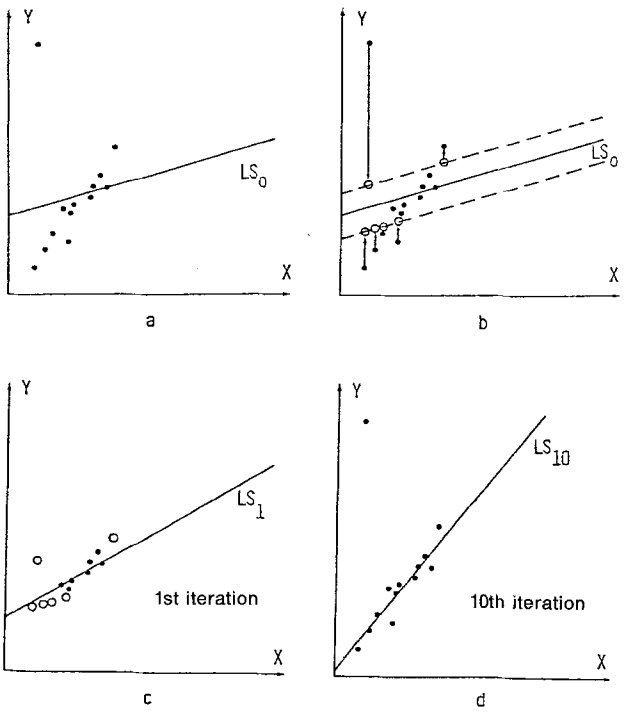

Figure 3. Computation of a Huber regression line.

gram 'c' shows the least squares line LS_1 which has been computed using the pseudo-observations. A new band is placed on LS_1, new pseudo-observations are computed, and so on. After 10 iterations one obtains the line LS_{10} indicated in diagram 'd'.

3.3 Testing of the linear hypotheses

In order to test a linear hypothesis expressed in the canonical form

$$H_0\colon\ \vartheta_{q+1} = \vartheta_{q+2} = \ldots = \vartheta_p = 0, \quad (0 \le q < p) \tag{3.5}$$

we use the following procedure (called the *τ-test*). Denote by ω the submodel of Ω obtained by imposing the condition H_0. Let $\hat{\vartheta}_{\Omega,1}, \ldots, \hat{\vartheta}_{\Omega,p}, \hat{\sigma}_\Omega$ be the solution of (3.3) and (3.4) and let $\hat{\vartheta}_{\omega,1}, \ldots, \hat{\vartheta}_{\omega,p}$ be a Huber estimate of $\vartheta_1, \ldots, \vartheta_q$ in ω, that is the solution of

$$\sum_i^n \psi((y_i - \sum_{j=1}^q \vartheta_j x_{ij})/\hat{\sigma}_\Omega) = 0, \quad j = 1, \ldots, q. \tag{3.6}$$

Notice that in (3.6) the scale parameter is set to $\hat{\sigma}_\Omega$. Denote by $r_{\Omega,1}, \ldots, r_{\Omega,n}$ the residuals under Ω and by $r_{\omega,1}, \ldots, r_{\omega,n}$ the residuals under ω. Then, the *τ-test statistic* is defined by:

$$F_\tau = \frac{2}{p-q} \sum_i^n (\varrho(r_{\omega,i}/\hat{\sigma}_\Omega) - \varrho(r_{\Omega,i}/\hat{\sigma}_\Omega)). \tag{3.7}$$

This reduces to the classical F-test statistic for $c \to \infty$. The null distribution of F_τ can be approximately evaluated as the distribution of a linear combination of independent chi-squared random variables.

3.4 Potency estimation

For the parallel line assay, assume that the submodel $Y = \mu V + \pi W + \gamma_1 L_1 + error$ of (2.8) has been retained. Obviously, block effects could be present. Let $\hat{\mu}$, $\hat{\pi}$ and $\hat{\gamma}_1$ be the Huber estimates of μ, π and γ_1. From the estimated regression lines $y_S = \hat{\mu} - \hat{\pi} + \hat{\gamma}_1 x_S$ and $y_T = \hat{\mu} + \hat{\pi} + \hat{\gamma}_1 x_T$ an estimate of $M = \log(\varrho)$ (the horizontal distance between the two straight lines) is easily derived as:

$$\hat{M} = (y_T - y_S)/\hat{\gamma}_1 = 2\hat{\pi}/\hat{\gamma}_1. \tag{3.8}$$

Clearly, $\hat{M}$ is a robust potency estimate in the dose-metameter scale and should be converted to the original scale in order to obtain an estimate $\hat{\varrho} = \exp(\hat{M})$ of ϱ.

Confidence limits for M can be found by Fieller's theorem, as in [1], Section 4.12, using the asymptotic normal distribution of the Huber estimates. Assume that $E(\hat{\pi}) = \pi$ and $E(\hat{\gamma}_1) = \gamma_1$, where $E(\cdot)$ denotes expected value. This property holds if the 'contaminating distribution' $g(y|x)$ in (2.6) is symmetric. Let $\hat{\sigma}^2(v_{ij})$, $i = 1,2$, $j = 1,2$ be an estimate of the covariance matrix of $2\hat{\pi}$ and $\hat{\gamma}_1$. The lower and upper limits M_L, M_U of an interval with confidence coefficient $1 - 2\alpha$ are

$$M_L = (A - B)/(1 - a) \text{ and } M_U = (A + B)/(1 - a), \tag{3.9}$$

where

$$A = \hat{M} - a v_{12}/v_{22}$$
$$B = z_\alpha \hat{\sigma}\{v_{11} - 2\hat{M}v_{12} + \hat{M}^2 v_{22} - a(v_{11} - v_{12}^2/v_{22})\}^{\frac{1}{2}}/\hat{\gamma}_1$$
$$a = z_\alpha^2 \hat{\sigma}^2 v_{22}/\hat{\gamma}_1^2 \tag{3.10}$$

and z_α is the upper α-quantile of the standard normal distribution. These limits can be converted to the original scale in order to obtain confidence limits for R.

For the slope ratio assay, assume that the submodel $Y = \mu V + \eta_1 X_S + \theta_1 X_T + error$ of (2.9) has been retained. Here, blanks are ignored; otherwise the term $\alpha_0 X_0$ is included. Again, block effects could be present. Let $\hat{\mu}$, $\hat{\eta}_1$ and $\hat{\theta}_1$ be the Huber estimates of μ, η_1 and θ_1. Then,

$$\hat{R} = \hat{\theta}_1/\hat{\eta}_1$$

is a robust estimate of ϱ^λ and $\hat{\varrho} = \hat{R}^{1/\lambda}$ a robust estimate of ϱ. Confidence limits for ϱ^λ can be computed as above. Assume that $E(\hat{\eta}_1) = \eta_1$ and $E(\hat{\theta}_1) = \theta_1$ and that an estimate of the covariance matrix of $\hat{\eta}_1$ and $\hat{\theta}_1$ of the form $\hat{\sigma}(v_{ij})$ is available. The lower and upper limits R_L, R_U of an interval with confidence coefficient $1 - 2\alpha$ are $R_L = (A - B)/(1 - a)$ and $R_U = (A + B)/(1 - a)$. Here A, B and a are obtained from (3.10), where $\hat{M}$ is replaced by $\hat{R}$ and $\hat{\gamma}_1$ by $\hat{\eta}_1$.

Remark. In (3.10) $\hat{\sigma}$ is the estimate of σ corresponding to the final submodel of (2.8) or (2.9). If the number of degrees of freedom ν in the residuals is relatively small (say, $\nu < 50$) it may be safer to use the upper α-quantile of the t-distribution with ν degrees of freedom in place of z_α.

4. Main feature of the statistical package ROBSYS

Although the use of robust statistical methods is frequently advocated and its theoretical aspects have received great attention during the past two decades, the most popular statistical packages are still very limited in this respect. Thanks to the support of the Swiss National Science Foundation two software facilities named ROBETH and ROBSYS [8,9] have been developed in order to fill this gap.

ROBETH is a FORTRAN subroutine library of numerical algorithms which allow the computation of many of the most recent procedures based on M-estimated, including robust regression, robust testing of linear hypotheses and robust covariances. This library is mainly meant for statisticians experienced in programming.

ROBSYS is a system control program which facilitates the use of ROBETH by avoiding the need of FORTRAN programming. It is based on a keyword (problem) oriented language of commands which allows a straightforward problem description stated in terms of the structure of the data and the kind of analysis and results required. ROBSYS is intended for the real data applications of a wide range of users.

Like many other packages, ROBSYS supplies the basic data manipulation tools such as access to data, definition of typical data structures (e.g. variables and cases), data transformation and selection, basic graphical and numerical displays. Furthermore, ROBSYS makes available the power of a FORTRAN based programming language by which the user can easily manipulate the complete data structure, including the results of the statistical modules.

Several special commands have been implemented in ROBSYS in order to perform robust statistical analysis of parallel line and slope ratio assays [10]. For example, the following program performs the usual tests and computes the potency estimate for a completely randomized unsymmetric design with 3 doses of S and 4 doses of T.

```
$DATa
 FILe is "EXMPLE.DAT"
 VARiables are "Y,DOSE,W"
$ADD "METAMETR"
$COMpute (Window to the FORTRAN compiler)
      METAMETR=ALOG(DOSE+.5)
$SET(change the c parameter default value)
      CPSI=1.5
$PARlin (def. the || line assay environment)
 TITLE "Example of parallel line assay"
 RESponse in "Y"; DOses in "METAMETR"
 PREparation (1/-1) in "W"
 SDOses = "3"; TDOses = "4"
$TESt "F-L"  (Deviation from linear fit)
$TESt "L-L1" (Parallelism)
$TESt "L1-T" (Linear term)
$TESt "P<C"  (Preparations)
```

ROBSYS is completely written in FORTRAN and enjoys the portability of this programming language. It has been successfully implemented on a large variety of machines, ranging from mainframes to microcomputers.

5. *An experimental scheme to validate and compare bioassay systems*

The statistical analysis of a parallel line or a slope ratio assay leads to a *valid* potency estimate under the fundamental *condition of similarity* (2.1). An important theoretical consequence is that *in an analytic dilution assay, the choice of the assay technique, in particular of the experimental material (cell or embryo culture or animal) should be irrelevant, although this may affect precision of the estimates.* In such an assay, the biological system plays a part analogous to that of a balance in weighing an object: it is an instrument, not a factor influencing the magnitude of the result (see Finney [1], Section 3.4).

It also follows that a *validation* of technique A (for example an alternative technique) with respect to a certain substance can be performed as follows:

I. Take two basic concentrations C_S and C_T of the substance. Consider the first concentration as the standard S and the second as the test preparation T. Notice that the ratio $\varrho = C_T/C_S$ is the potency of T with respect to S; furthermore, this ratio is known and the condition of similarity is clearly satisfied.

II. Perform a biological assay using the technique A and several (derived) concentrations of S and T.

III. Compute (with the help of ROBSYS) an estimate $\hat{\varrho}_A$ of ϱ as well as an estimate of its precision (for example its 95% confidence interval).

If the assay is well conducted, we expect $\hat{\varrho}_A$ to be very close to ϱ (*validation*). Finally, in order to *compare* technique A to technique B it suffices to:

IV. Repeat the same assay using technique B and obtain an estimate $\hat{\varrho}_B$ of ϱ.

The comparison between A and B is then based essentially on the comparisons of the *statistical precisions* of $\hat{\varrho}_A$ and $\hat{\varrho}_B$ and of the possible *absolute biases* $|\varrho - \hat{\varrho}_A|$ and $|\varrho - \hat{\varrho}_B|$.

6. *Experimental results and statistical analysis*

The experimental scheme described in Section 5 was applied to three alternative biological assay systems [3,4,6,7,11,12]. The experimental subjects were respectively: cultured chicken embryos, cultured rodent embryos and samples of an aggregating brain cell culture from fetal rats. Two substances were used: diphenylhydantoin (DPH) and caffein (CAF). In each experiment (with one exception), subjects were randomly allocated to 8 non-zero doses (4 of S and 4 of T) and several responses were measured for each subject.

General remarks

For each dose response relation, a preliminary exploratory data analysis was performed in order to determine the appropriate types of analysis (parallel line–PL–or slope ratio–SR–assay), response metameters ($Y = H(U)$) and dose ranges.

In tables 1–6, column 'U' indicates the abbreviated names of the responses; column 'Y' indicates the response metameter and column 'Type' the type of analysis. The p-values for the test of linearity are given in column 'H_L'; those for the test of parallelism (in the PL case) or the test of intersection at the origin (in the SR case) are given in column '$H_{P/T}$'. In the following columns, the reader can find the estimates of η_1 (the slope of the S line, i.e. the common slope γ_1 in the PL case), σ (the standard error), ν (the number of degrees of freedom in $\hat{\sigma}$), and ϱ (the potency in the original dose scale). Confidence intervals for ϱ with coefficient 95% are given in column (R_L, R_U). For each response, computations were done according to the classical procedure (first line) and to the robust procedure described in Section 3 with $c = 1.35$ (second line). Comments on the last two columns can be found in Section 6.7.

Table 1. Assay of DPH using chicken embryos ($\varrho = 1.50$)

U	Y	Type	H_L	$H_{P/T}$	$\hat{\eta}_1$	$\hat{\sigma}$	v	$\hat{\varrho}$	(R_L, R_U)	$\lvert\frac{\hat{\sigma}}{\hat{\eta}_1}\rvert$	a	$4n$
SN	U	PL	0.58	0.22	−0.77	1.35	40	1.66	(1.17, 2.35)	1.74	0.070	69
		with blocs	0.65	0.19	−0.78	1.28	40	1.62	(1.14, 2.28)	1.63	0.070	64
EL	$\sqrt{U - U_0}$*	PL	0.78	0.20	−0.07	0.33	36	1.54	–	4.86	1.215	493
		no blocs	0.77	0.24	−0.09	0.28	36	1.50	(0.61, 4.13)	2.99	0.477	227
AP	$U^2/100$	PL	0.20	0.99	−0.65	3.27	40	1.53	(0.28, 5.82)	5.03	0.580	548
		with blocs	0.34	0.69	−0.78	3.26	40	1.37	(0.43, 3.29)	4.18	0.380	398
AV	U	PL	0.83	0.24	−2.00	4.05	37	1.55	(0.94, 2.34)	2.03	0.132	92
		with blocs	0.90	0.13	−2.06	3.33	37	1.59	(1.08, 2.26)	1.61	0.092	62
HW	U	PL	0.70	0.71	1.44	4.75	27	1.54	(0.36, 10.2)	3.29	0.738	237
		no blocs	0.69	0.76	1.28	3.69	27	1.37	(0.51, 3.47)	2.89	0.490	191
HL	$\sqrt{U}$	PL	0.07	0.50	0.24	0.57	24	1.30	(0.29, 2.71)	2.39	0.573	125
		no blocs	0.04	0.55	0.24	0.51	24	1.36	(0.42, 2.68)	2.16	0.498	106

* U_0 is the embryo length at the beginning of the experience.

Table 2. Assay of CAF using chicken embryos ($\varrho = 1.41$)

U	Y	Type	H_L	$H_{P/T}$	$\hat{\eta}_1$	$\hat{\sigma}$	v	$\hat{\varrho}$	(R_L, R_U)	$\lvert\frac{\hat{\sigma}}{\hat{\eta}_1}\rvert$	a	$4n$
SN	U	PL	0.70	0.69	−2.09	3.60	37	1.10	(0.62, 1.60)	1.73	0.179	64
		no blocs	0.96	0.77	−2.24	2.17	37	1.30	(1.01, 1.63)	0.97	0.063	23
EL	$\sqrt{U}$	PL	0.92	0.31	−0.12	0.19	43	1.19	(0.82, 1.63)	1.54	0.106	55
		no blocs	0.81	0.86	−0.13	0.15	43	1.24	(0.93, 1.60)	1.22	0.068	32
AP	U	PL	0.03	0.57	−3.28	8.54	46	1.63	(0.99, 2.79)	2.61	0.113	140
		with blocs	0.06	0.53	−3.49	6.89	46	1.77	(1.18, 2.77)	1.98	0.077	86
AP	U	PL	0.81	0.36	−7.07	8.37	43	1.21	(0.93, 1.55)	1.18	0.062	32
		no blocs	0.55	0.46	−7.06	6.34	43	1.33	(1.09, 1.62)	0.90	0.039	20
AV	$\sqrt{U}$	PL	0.76	0.63	−0.50	1.16	53	1.36	(0.90, 2.16)	2.33	0.065	111
		with blocs	0.88	0.70	−0.49	1.16	53	1.46	(0.95, 2.38)	2.37	0.070	122
AV	$\sqrt{U}$	PL	0.35	0.11	−0.89	1.16	43	1.06	(0.77, 1.39)	1.30	0.076	38
		no blocs	0.19	0.32	−0.91	1.28	43	1.11	(0.80, 1.47)	1.41	0.081	46

Table 3. Assay of DPH using rodent embryos ($\varrho = 1.20$)

U	Y	Type	H_L	$H_{P/T}$	$\hat{\eta}_1$	$\hat{\sigma}$	v	$\hat{\varrho}$	(R_L, R_U)	$\lvert\frac{\hat{\sigma}}{\hat{\eta}_1}\rvert$	a	$4n$
SN	U	SR	0.75	0.50	−3.91	0.83	38	1.63	(1.35, 2.02)	0.21	0.046	33
			0.68	0.40	−3.90	0.82	38	1.58	(1.31, 1.97)	0.21	0.047	34
HL	U	SR	0.21	0.85	−0.16	0.06	44	1.86	(1.35, 2.88)	0.40	0.145	85
			0.21	0.78	−0.16	0.07	44	1.83	(1.33, 2.82)	0.43	0.144	121
YS	U	SR	0.35	0.73	−0.35	0.12	44	1.56	(1.16, 2.23)	0.33	0.106	72
			0.58	0.70	−0.36	0.11	44	1.54	(1.18, 2.13)	0.32	0.087	60
CR	$\sqrt{U}$	SR	0.07	0.57	−0.11	0.03	44	1.65	(1.31, 2.17)	0.26	0.066	46
			0.13	0.52	−0.12	0.03	44	1.65	(1.31, 2.20)	0.28	0.071	41
AI	$\sqrt{U}$	SR	–	–	2.60	0.66	44	1.14	–	0.25	0.059	–
			–	–	2.60	0.73	44	1.14	–	0.28	0.083	–

Table 4. Assay of CAF using rodent embryos ($\varrho = 1.11$)

U	Y	Type	H_L	$H_{P/T}$	$\hat{\eta}_1$	$\hat{\sigma}$	v	$\hat{\varrho}$	(R_L, R_U)	$\lvert\frac{\hat{\sigma}}{\hat{\eta}_1}\rvert$	a	$4n$
SN	U	PL	0.19	0.76	−0.95	1.95	38	1.13	(1.00, 1.29)	2.06	0.103	11
			0.10	0.62	−0.95	1.97	38	1.16	(1.03, 1.32)	2.08	0.098	12
HL	U	PL	0.18	0.41	−0.05	0.14	44	1.13	(0.97, 1.36)	2.80	0.132	17
			0.11	0.40	−0.05	0.12	44	1.14	(1.01, 1.32)	2.22	0.087	14
YS	U	PL	0.07	0.63	−0.08	0.27	38	1.33	(1.11, 1.79)	3.15	0.241	24
			0.02	0.52	−0.08	0.23	38	1.31	(1.12, 1.66)	2.68	0.180	19
CR	$\sqrt{U}$	PL	0.54	0.88	−0.03	0.06	38	1.18	(1.07, 1.31)	1.68	0.068	11
			0.73	0.84	−0.03	0.06	38	1.19	(1.08, 1.32)	1.76	0.068	11
AI	$\sqrt{U}$	PL	0.20	0.24	0.26	0.32	44	1.08	(1.02, 1.16)	1.21	0.024	7
			0.22	0.09	0.27	0.31	44	1.10	(1.02, 1.18)	1.17	0.026	6

Table 5. Assay of DPH using an aggregating cell culture ($\varrho = 1.50$)

U	Y	Type	H_L	$H_{P/T}$	$\hat{\eta}_1$	$\hat{\sigma}$	ν	$\hat{\varrho}$	(R_L, R_U)	$\lvert\frac{\hat{\sigma}}{\hat{\eta}_1}\rvert$	a	$4n$
PROT	$\sqrt{U}$	SR	0.44	0.19	−0.74	0.08	31	1.79	(1.61, 2.03)	0.10	0.017	11
			0.55	0.21	−0.72	0.08	31	1.82	(1.63, 2.09)	0.11	0.020	11
DNA	$\frac{U^2}{100\,000}$	PL	0.13	0.75	−0.97	0.54	20	1.64	(1.38, 1.99)	0.56	0.021	10
			0.14	0.40	−1.00	0.50	20	1.65	(1.41, 1.97)	0.51	0.018	9
CAT	$\frac{U^2}{10\,000}$	PL	0.23	0.41	−0.26	0.23	27	2.35	(1.85, 3.10)	0.88	0.022	20
			0.14	0.50	−0.25	0.23	27	2.31	(1.84, 3.00)	0.92	0.020	22
CNP	U^2	PL	0.21	0.79	−1.55	0.91	16	1.52	(1.25, 1.87)	0.58	0.039	11
			0.54	0.63	−1.56	0.64	16	1.47	(1.28, 1.69)	0.41	0.019	7
GS	$\frac{U}{100\,000}$	SR	0.24	0.86	−1.18	0.11	31	1.51	(1.36, 1.69)	0.10	0.014	9
			0.11	0.98	−1.19	0.11	31	1.51	(1.37, 1.68)	0.09	0.013	10
GAD	$\frac{U}{1000}$	PL	0.12	0.19	−0.32	0.28	28	1.67	(1.34, 2.13)	0.87	0.020	19
			0.15	0.21	−0.32	0.29	28	1.69	(1.34, 2.17)	0.90	0.020	21

Table 6. Assay of CAF using an aggregating cell culture

U	Y	Type	H_L	$H_{P/T}$	$\hat{\eta}_1$	$\hat{\sigma}$	ν	$\hat{\varrho}$	(R_L, R_U)	$\lvert\frac{\hat{\sigma}}{\hat{\eta}_1}\rvert$	a	$4n$
PROT	$\sqrt{U}$	SR	0.25	–	−0.73	0.04	12	–	–	0.06	–	6
			0.21	–	−0.73	0.05	12	–	–	0.06	–	6
DNA	$\frac{U}{100}$	SR	0.25	–	−3.09	0.39	12	–	–	0.13	–	12
			0.29	–	−3.13	0.41	12	–	–	0.13	–	13
CNP	U	SR	0.97	–	−0.98	0.15	12	–	–	0.15	–	16
			0.97	-	−1.01	0.14	12	–	–	0.14	–	14
GS	$\frac{U}{10\,000}$	SR	0.82	–	−6.25	0.69	12	–	–	0.11	–	10
			0.73	–	−6.28	0.65	12	–	–	0.10	–	9
GAD	$\frac{U^2}{1\,000\,000}$	SR	0.95	–	−4.83	2.05	12	–	–	0.42	–	101
			0.99	–	−4.12	1.62	12	–	–	0.39	–	91

Figures 4–9 show the dose response diagrams in metametric scales. Doses are labelled by B, 1, 2, 3,..., with B representing the zero dose. The original dose values are given in the following subsections. Unfortunately, in some cases, appropriate preliminary investigations to determine the form of the dose-response function F could not be performed and some non-monotonic relationships were observed. In order to use the standard models of PL and SR assays, attention was restricted to monotonic segments for which a linearizing transformation could be found. For this reason, some data could not be used in the final linear fits. The largest dose ranges where linearity could be accepted were determined. These domains are marked by bold line segments on the x-axis of the diagrams, and the doses used in the estimation of η_1, σ and ϱ are underlined. The figures include also the robust linear fits for S (continuous line) and T (broken line). The points are adjusted for significant block effects.

For PL assays, the following dose scale transformation is used in the diagrams and in the expression of $\hat{\eta}_1$. Let $z_{Sh} = D^{(h-1)}C_S$, $z_{Th} = D^{(h-1)}C_T$, $h = 1, 2, 3, 4$, be the original doses of S and T. Here, C_S and C_T are the two lowest concentrations and D is a constant factor. We set $x_{Sh} = 2\ln(z_{Sh}/C_S)/\ln(D) + 1$, $x_{Th} = 2\ln(z_{Th}/C_T)/\ln(D) + 1$, $h = 1, 2, 3, 4$ and notice that these values correspond approximatively to the labels 1, 2, 3,...,8 on the diagrams. In order to obtain $\hat{\eta}_1$ in the conventional log-scale the values given in the tables should be multiplied by $2/\ln(D)$.

In SR assays, the dose scales of S and T were adjusted so that the highest dose of each preparation is unity. In order to obtain $\hat{\eta}_1$ in the conventional scale the values given in the tables should be divided by the highest dose of S.

6.1 Assay of DPH *using chicken embryos*

Responses: embryonic length (*EL*, div_1), area pellucida diameter (*AP*, div_1), area vasculosa diameter (*AV*, div_1), width of the heart (*HW*, div_2), length of the heart (*HL*, div_2), somite number (*SN*). (The units 'div_1' and 'div_2' are division numbers on the microscopic image: 1 mm = 5.8 div_1 = 25 div_2.)

Doses of S (μg/ml): 14 (1), 28 (3), 56 (5), 112 (7).

Doses of T (μg/ml): 21 (2), 42 (4), 84 (6), 168 (8).

True potency: $\varrho = 1.5$; *Concentration factor*: $D = 2$; *Blanks*: present.

Subject number: 8 per dose (however, some of the response values are missing).

Design: symmetric randomized blocks (9 subjects were allocated to 9 doses on 8 different days = blocks).

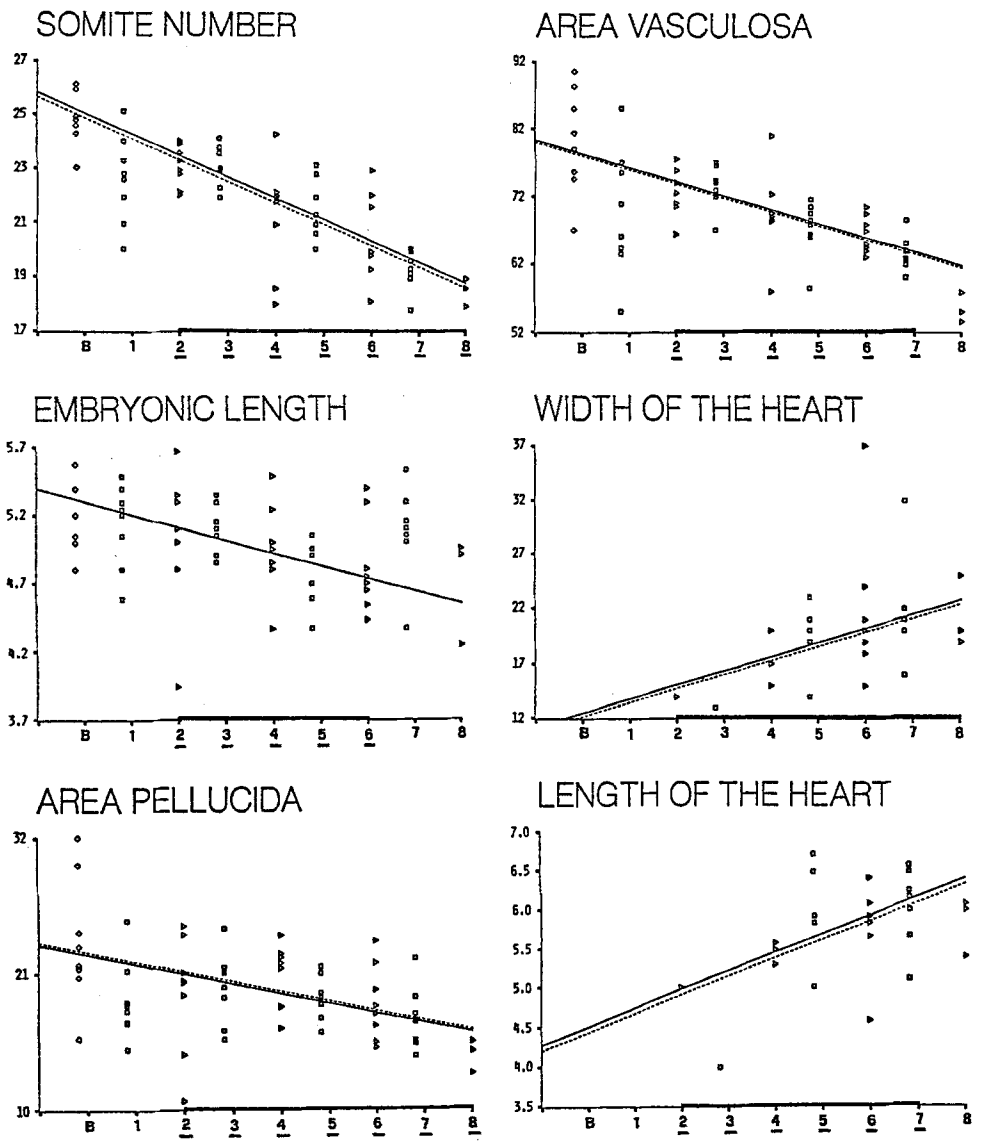

Figure 4. Assay of DPH using chicken embryos.

Observations were taken after 17 h, 24 h and 42 h of incubation. The results at 42 h seem the most appropriate for potency estimation: in fact, the dose response relations at 17 h and 24 h are very weak for all responses and the potency estimates very imprecise. For this reason, only the results at 42 h are presented in table 1 and figure 4. However, subsequent experience has shown that steeper relations can be observed at 66 h of incubation.

Block effects are retained in all final models except for *EL*, *HL* and *HW* (tests were performed at the 0.07 level). Potency estimates (particularly robust estimates) are very close to the true value 1.5. However, robust values of $\hat{\sigma}$ are lower than classical values. This can be explained by the presence of some mild outliers. Despite the huge confidence interval, the robust estimate of ϱ for the response *EL* is 1.5 exactly. The classical procedure gives 1.54, but the confidence interval cannot be computed according to (3.9)–(3.10) because the slope γ_1 is very small and a is larger than 1. Also notice that, for *EL*, the hypothesis H_R: $\gamma_1 = 0$ is rejected by the robust procedure at the level 0.003, whereas the p-value of the classical F-test is 0.07.

The analysis of *AP* without metametric transformation (i.e. with $Y = U$) produces essentially the same results as the analysis with $Y = U^2/100$ ($\hat{\varrho} = 1.66\backslash 1.46$; $\hat{\eta}_1 = -0.80\backslash -0.96$; $\hat{\sigma} = 3.93\backslash 3.87$).

Doses lower than 21 µg/ml are excluded from all linearity domains. However, dose 1 could be included in the domain of *AP* with $Y = U^2/100$ (H_L and H_P accepted at the respective levels – classical\robust – of 0.17\0.28 and 0.34\0.10) and with $Y = U$ (H_L and H_P accepted at the respective levels of 0.14\0.22 and 0.30\0.08) but this would impair the value of $\hat{\varrho}$ (1.05\0.89 and 1.20\0.99).

6.2 Assay of CAF *using chicken embryos*

Responses: embryonic length (*EL*, mm), area pellucida (*AP*, mm^2), area vasculosa (*AV*, mm^2), somite number (*SN*).

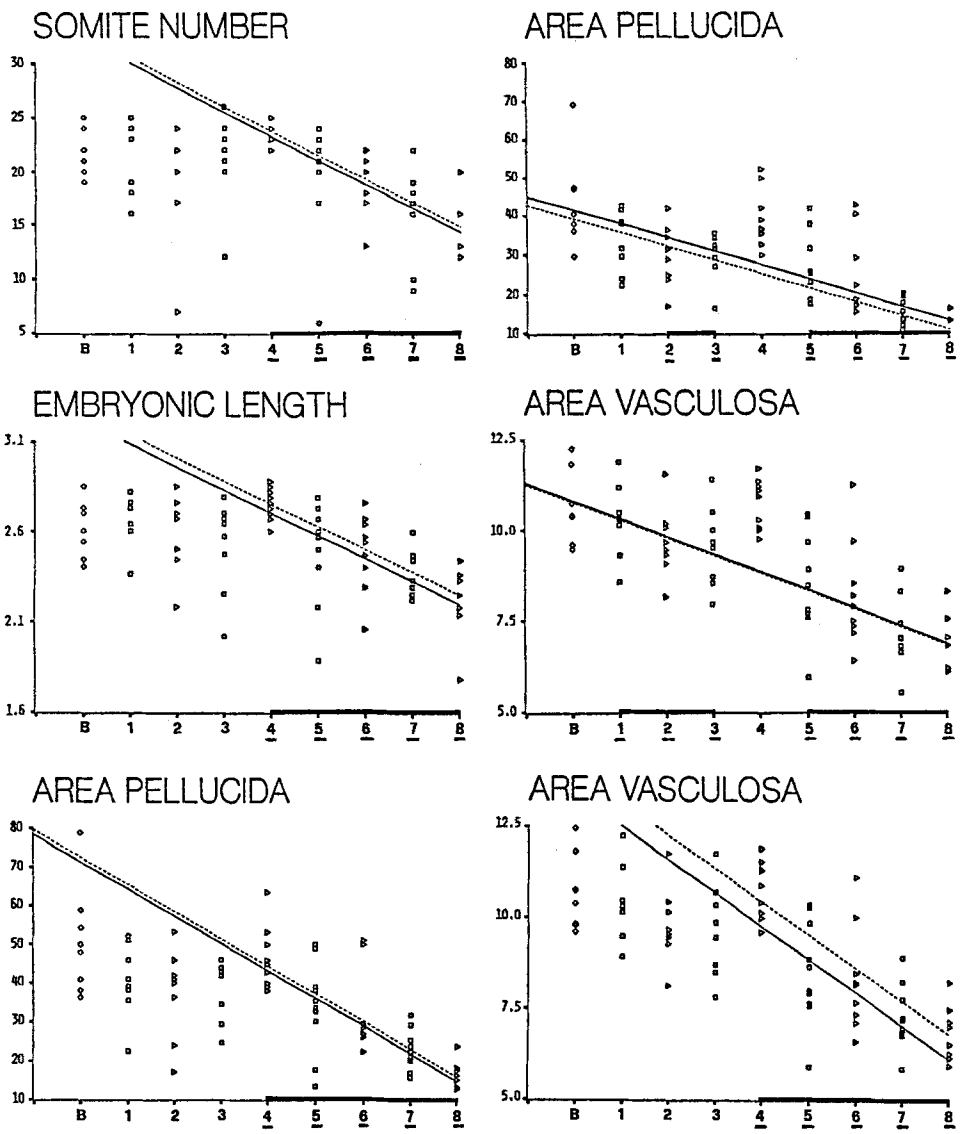

Figure 5. Assay of CAF using chicken embryos.

Doses of S (µg/ml): 73.72 (1), 143.56 (3), 283.24 (5), 554.84 (7).
Doses of T (µg/ml): 102.82 (2), 201.76 (4), 395.76 (6), 776.00 (8).
True potency: $\varrho = 1.41$; *Concentration factor*: $D \approx 2$; *Blanks:* present.
Subject number: 10 per dose (however, some of the response values are missing).
Design: symmetric randomized blocks (10 subjects were allocated to 9 doses on 5 different days = blocks).
Observations were taken after 42 h of incubation. Several mild outliers can be observed on the dose-response diagrams (see *SN* in particular). This explains the lower values of $\hat{\sigma}$ obtained with the robust procedure.
In all diagrams, dose 4 produces relatively high response values and seems to separate the dose-response relations into two parts. This could be related to an exciting effect of CAF at low doses. Specifically, for *EL* and *SN*, doses 1, 2 and 3 must be excluded from the final models. For *AP* and *AV*, two linear fits can be accepted: the first one excludes dose 4 and retains the block effects, the second one excludes doses lower than 4 and does not include blocks. The first model produces a smaller bias; the second one a higher precision of $\hat{\varrho}$. More data are necessary in order to decide between these two models.
Interestingly, *EL* and *AP* are among the worst responses to DPH and among the best to CAF.

6.3 Assay of DPH *using rodent embryos*
Responses: yolk sac (*YS*, mm), crown-rump length (*CR*, mm), head length (*HL*, mm), somite number (*SN*), incidence of abnormalities (*AI*).
Doses of S (µg/ml): 8.0 (1), 20 (3), 50 (5), 125 (7).
Doses of T (µg/ml): 9.6 (2), 24 (4), 60 (6), 150 (8).
True potency: $\varrho = 1.2$; *Concentration factor*: $D = 2.5$; *Blanks*: absent.
Subject number: 6 per dose.
*Design:*symmetric randomized.

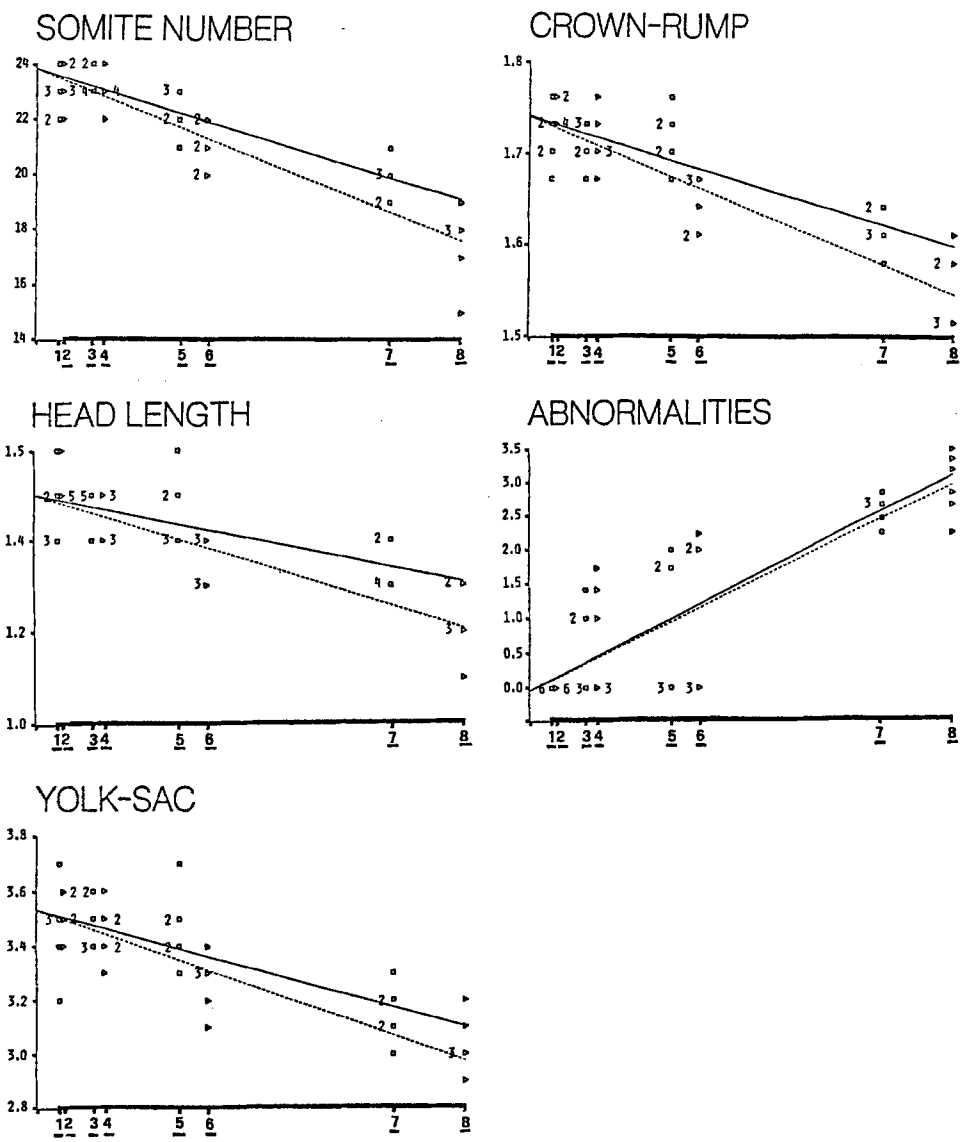

Figure 6. Assay of DPH using rodent embryos.

Observations were taken after 48 h of incubation. In the diagrams, several response values are represented by the same points and their frequencies are indicated.

On the basis of the statistical tests, the SR assay model is generally accepted on the complete dose range. However, diagram inspection suggests that an appreciable reaction to DPH is not obtained for doses lower than 50 µg/ml. (Interestingly, chicken embryos react to DPH for doses higher than 21 µg/ml.) If attention is restricted to doses 5, 6, 7 and 8, the dose-response regressions of T are generally lower than the S ones. For this reason, most estimates of ϱ reported in table 3 are significantly higher than 1.2. (This bias could be due to a small error in the doses. Doses were difficult to prepare due to the narrow concentration range.)

As the number of abnormalities at low doses is often zero, the distribution of AI for given z is clearly non-Gaussian, and an inference based on (2.5) or even (2.6) would hardly be appropriate. For this reason, p-values and confidence intervals are not given in table 3. Yet, $\hat{\varrho}$ is very close to 1.2.

6.4 Assay of CAF *using rodent embryos*

Responses: yolk sac (*YS*, mm), crown-rump length (*CR*, mm), head length (*HL*, mm), somite number (*SN*), incidence of abnormalities (*AI*).

Doses of S (µg/ml): 208 (1), 250 (3), 300 (5), 360 (7).

Doses of T (µg/ml): 231 (2), 277 (4), 333 (6), 400 (8).

True potency: $\varrho = 1.11$; *Concentration factor*: $D \approx 1.2$;

Blanks: absent.

Subject number: 6 per dose.

Design: symmetric randomized.

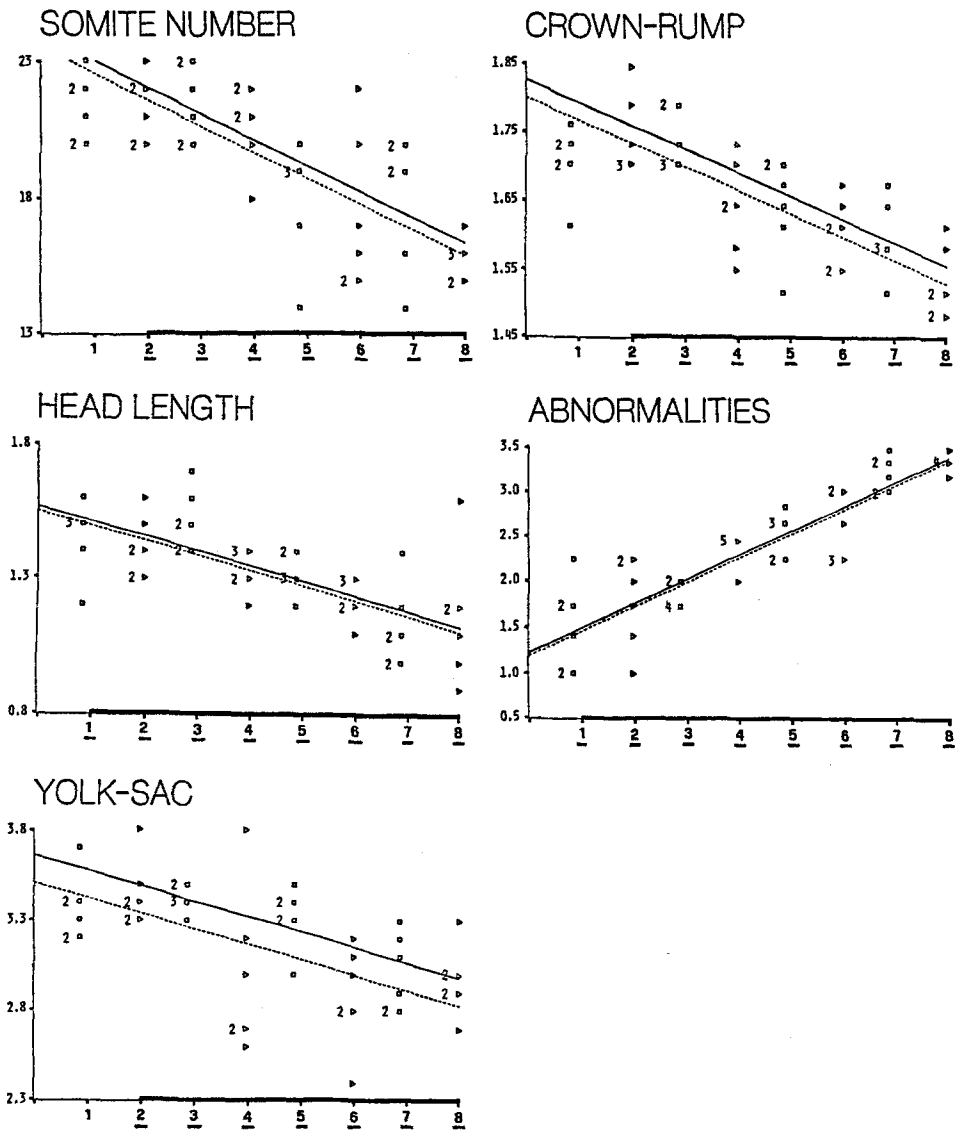

Figure 7. Assay of CAF using rodent embryos.

Observations were taken after 48 h of incubation. On the diagrams, several response values are represented by the same points and their frequencies are indicated.

The diagrams of *HL*, *SN*, *YS* and *CR* suggest a non-decreasing pattern to the left of dose 2 or 3. Again, this could be related to a possible exciting effect of CAF, but this hypothesis should be confirmed with supplementary data. (For comparison, the highest responses of chicken embryos to CAF are obtained at about 200 µg/ml.)

We observe that the robust values of $\hat{\sigma}$ are generally lower than the classical values.

6.5 Assay of DPH *using an aggregating cell culture*

Responses: protein (*PROT*, u_1), DNA, (*DNA*, u_1), glutamic acid decarboxylase activity (*GAD*, u_2), glutamine synthetase activity (*GS*, u_2), choline acetyltransferase activity (*CAT*, u_2), 2′,3′-cyclic nucleotide 3′-phosphohydrolase activity (*CNP*, u_3). Here, u_1 = µg, u_2 = pmole/min/mg-*PROT*, u_3 = µmole/min/mg-*PROT*.

The specific enzyme activities per minute and mg of protein (mg-*PROT*) were observed. However, statistical analyses concern enzyme activities per test-culture. These are the products of the specific enzyme activities by the quantity of protein in the culture. In other words, the responses *GS*, *GAD*, *CNP* and *CAT* are recoded to *GS* · *PROT*, *GAD* · *PROT*, *CNP* · *PROT* and *CAT* · *PROT*.

Doses of S (µg/ml): 10 (1), 20 (3), 40 (5), 80 (7).

Doses of T (µg/ml): 15 (2), 30 (4), 60 (6), 120 (8).

True potency: $\varrho = 1.5$; *Concentration factor*: $D = 2$; *Blanks*: present.

Subject number: 4 per dose (one response is missing).

Design: symmetric randomized.

The estimates of ϱ obtained from *GS*, *GAD*, *CNP* and *DNA* are very close to the true value; some are also very precise (see *GS* in particular).

On the contrary, *PROT* and *CAT* produce biased estimates of ϱ. Notice that *PROT* is a multiplicative factor in

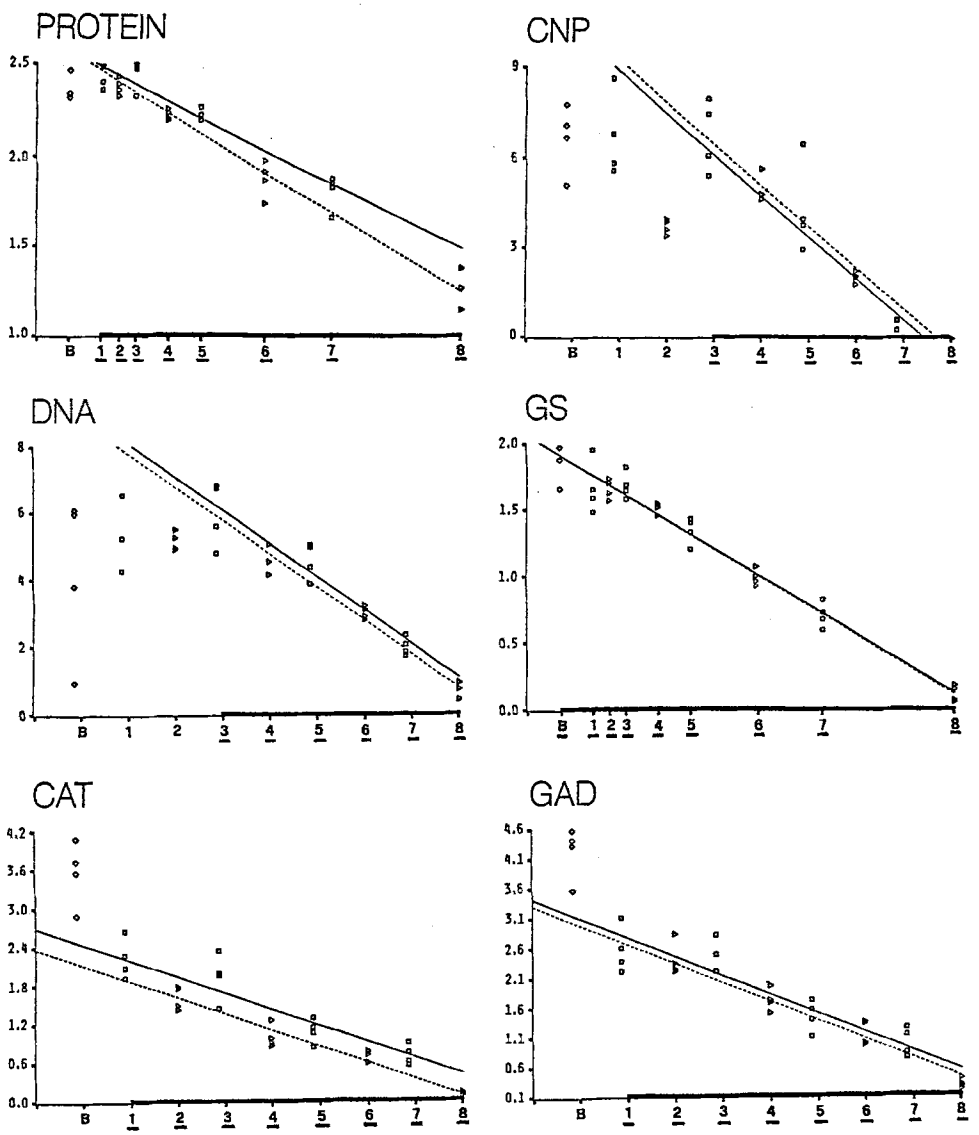

Figure 8. Assay of DPH using an aggregating cell culture.

GS, *GAD*, *CNP* and *CAT*. Yet, the bias of *PROT* is not reflected in *GS*, *GAD* and *CNP*. Also notice that an analysis of the original (non recoded) values of *CAT* versus 'dose' does not indicate any clear relationship. In other words, the dose-response diagram of *CAT* is probably determined by the dose-response pattern of *PROT*.

6.6 Assay of CAF *using an aggregating cell culture*
Responses: protein (*PROT*, u_1), DNA(*DNA*, u_1), glutamic acid decarboxylase activity (*GAD*, u_2), glutamine synthetase activity (*GS*, u_2), choline acetyltransferase activity (*CAT*, u_2), 2′,3′-cyclic nucleotide 3′-phosphohydrolase activity (*CNP*, u_3).
The units 'u_1', 'u_2' and 'u_3' are defined as in 6.5; the responses *GS*, *CAT*, *GAD* and *CNP* are recoded as in 6.5.
Doses of S (μg/ml): 0 (B), 30 (1), 100 (2), 300 (3), 600 (4), 900 (5).
Doses of T: none.
True potency: none; *Blanks*: present.
Subject number: 3 per dose.
Design: the design of this experiment does not follow the scheme of Section 5: in fact, no distinction has been made between doses of *T* and doses of *S*. Subjects were randomly allocated to 6 different doses.

There is not potency to be estimated. However, the straight line model can be accepted for all dose-response relationships, except for *CAT*. This means that the SR assay type of analysis would be appropriate. Classical and robust estimates of η_1 (the slope) and σ are generally very similar, except for the response *GAD* which contains some outliers.
Linearity domains include all doses. However, we suspect an exciting effect of CAF for concentrations lower than 100 μg/ml. If confirmed, the dose range should be conveniently reduced.

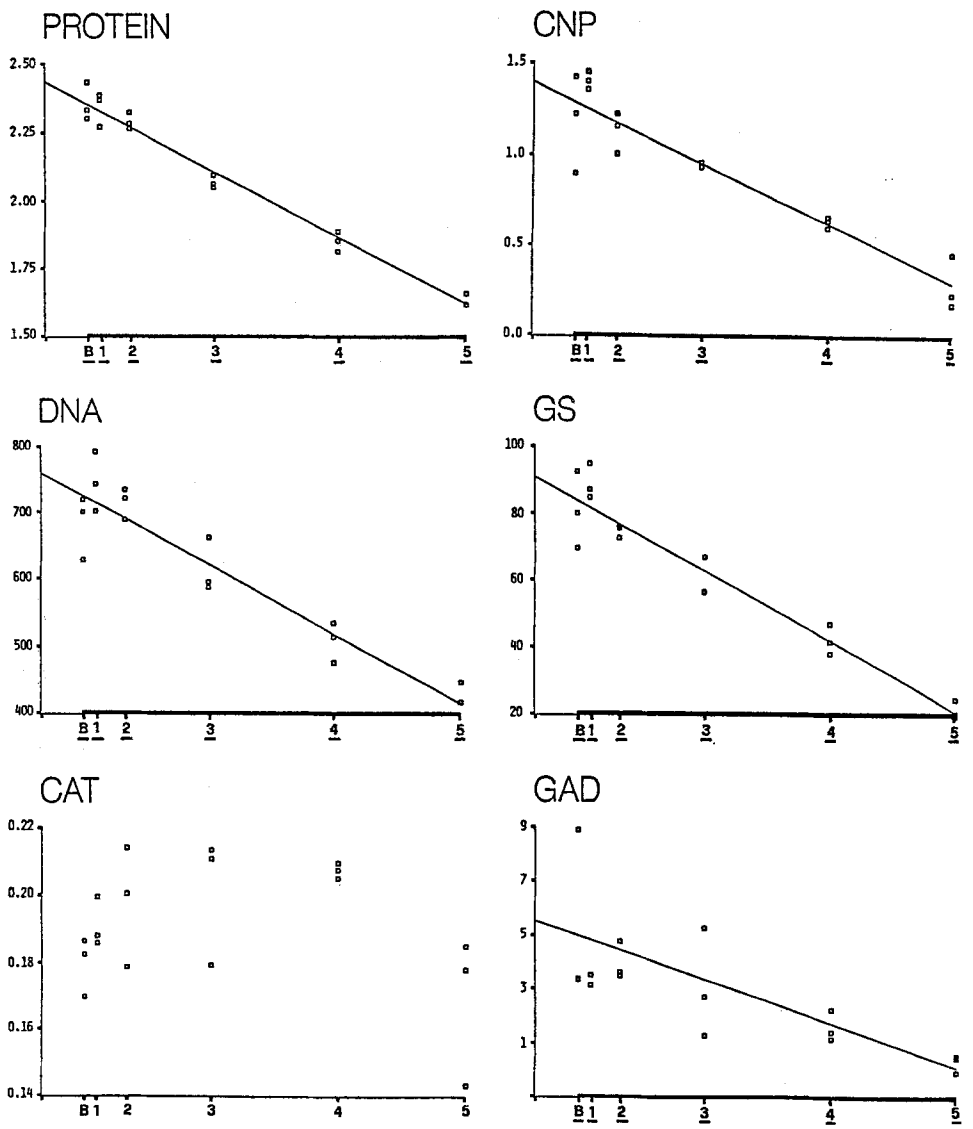

Figure 9. Assay of CAF using an aggregating cell culture.

6.7 Statistical comparisons

As mentioned in Section 5, the statistical comparison between assay techniques can be based essentially on their capability of producing potency estimates to a certain degree of accuracy. For a given dose-response relation, the estimate accuracy is measured, for example, by the absolute bias $|\varrho - \hat{\varrho}|$ and the length $|R_U - R_L|$ of the confidence interval for ϱ. The latter depends on the number of observations used in the estimation of the final dose-response model. Unfortunately, for many assays considered in this paper, several observations could not be taken into account because they fell outside the domains of linearity. Therefore, in order to eliminate the effect of different numbers of observations in the comparison of the confidence intervals, an alternative procedure is proposed.

Suppose that we need to plan an experiment for estimating the potency of a new test preparation T with respect to DPH or CAF (S in the following), using one of the systems considered in 6.1–6.6 and a specific response. Suppose also, that an advance estimate of ϱ (available from a collateral source of information) is $R_0 = 1$; (notice that this is no loss of generality: if $R_0 \neq 1$, T or S can be appropriately diluted). Finally, assume that a symmetric optimal design (see Finney[1], Sections 6.8 and 8.1) with 2 doses of each preparation and n observations at each dose has to be used. The doses will be chosen at the extremes, x_L and x_U, of the domains of linearity, which coincide for T and S, and can be determined with the help of figures 4–9.

For a PL assay it can be assumed (without loss of generality) that $-x_L = x_U$. As we expect $\hat{M}$ to be close to 0, it follows from (3.9)–(3.10) that

$$M_L, M_U \approx \pm t_{\alpha,\nu}\left(\frac{\sigma}{\eta_1}\right)\{f(1-a)/n\}^{\frac{1}{2}}/(1-a)$$

with $a = t^2_{\alpha,\nu}(\sigma/\eta_1)^2 f/(4nx_U^2)$. Here, $t_{\alpha,\nu}$ is the upper α-quantile of the t-distribution with $\nu = 4n-3$ degrees of freedom.

For a SR assay with n observations of both preparations at x_L and x_U, one obtains from (3.9)–(3.10), using $\hat{R} = 1$,

$$R_L, R_U \approx \left(1 \pm t_{\alpha,\nu}\left(\frac{\sigma}{\eta_1}\right)\{fv_{22} + fv_{11}(1-a)\}^{\frac{1}{2}}\right)/(1-a)$$

with $a = t^2_{\alpha,\nu}(\sigma/\eta_1)^2 fv_{22}$. Here, $(v_{ij}), i,j = 1,2,3$ is the inverse of the matrix (u_{ij}) such that $u_{11} = u_{22} = n(x_L^2 + x_U^2), u_{33} = 4n, u_{12} = 0, u_{13} = u_{23} = n(x_L + x_U)$. By fixing η_1 and σ according to the entries of tables 1–6 and setting $\alpha = 5\%$, $c = 1.35$, we can determine n so that the relative uncertainty $(R_U - R_L)/\varrho$ (with $\varrho = 1$) equals a specified value, say 60%. The results of these computations are shown in the last column of tables 1–6.

Remark. In the PL case, the value of $\hat{\sigma}/\hat{\eta}_1$ (exhibited in the tables) may be regarded as an effective standard deviation per response in respect of the measure of log potency. Therefore, alternative techniques for the estimation of the potency may be compared in terms of $\hat{\sigma}/\hat{\eta}_1$ (see Finney[1], Section 6.11). For SR assays with standardized S scale (i.e. such that the highest dose is 1), $\hat{\sigma}/\hat{\eta}_1$ plays a similar role, though it cannot be given a simple interpretation (see Finney[1], Section 8.6). However, no simple comparison between PL and SR assays can be made on these grounds.

7. *Discussion and conclusions*

As expected, the results of classical and robust analyses are similar in most situations. No data set contains highly deviant observations capable of distorting coefficient and potency estimates an appreciable amount.

However, a moderate number of mild outliers inflates many of the classical estimates of the error variance. As a consequence, on a few occasions, the classical F-test and the robust τ-test led to opposite decisions (see Section 6.1). Moreover, this variance inflation is reflected in an increased number of observations being necessary to obtain a specified precision of the potency estimate. For some responses (see the assays using chicken embryos), the robust procedure is likely to require about 2/3 (or even 1/2!) the number of observations which would be required by the classical procedure.

The experiments evaluated in this paper were performed with the high level of attention to detail appropriate to a research project. Nevertheless, a small number of anomalous observations could not be avoided. It could be argued that, in routine experiments, the need of robust procedures would be correspondingly greater.

Several quantitative dose-response relationships were evaluated with regard to their capability of comparing a test preparation to caffein and diphenylhydantoin. Unfortunately, in some cases, appropriate pilot investigations to determine the form of the regression function could not be performed. For this reason, the statistical analysis was restricted to data on monotonic dose-response segments corresponding to toxic concentrations. (Interestingly, the toxicity domains of the two embryo based systems reveal a remarkable overlap!) Nevertheless, the results presented in Section 6 allow one to determine the most reliable responses for the underlying potency estimation problem.

In general, the best responses of the three systems can produce the desired results to various degrees of bias and precision. The estimates obtained with chicken embryos are the least biased; nevertheless, their standard errors are relatively high. The results of the assay of caffein using rodent embryos are very satisfactory; on the other hand, the most biased (but quite precise) estimates are obtained from the assay of DPH using the same system. (However, the appropriateness of the statistical model used in the analysis of this assay should still be investigated with the help of supplementary data). Despite the impossibility of investigating a potential bias in the assay of caffein, the sytem based on aggregating cell culture seems to produce the most stable responses and results. Homogeneity of the experimental material could be related to the different response variabilities observed in the three systems. The most similar experimental subjects are clearly obtained from the same cell culture, whereas the origin, the stocking time, and the explantation age of chicken embryos, are likely to be important sources of variation.

It must be stressed that no comparison or conclusion can be drawn beyond the experimental settings, the substances and the responses investigated. In particular, we did not analyse quantal responses, such as the frequency of anomalous embryos. Nor, did we attempt to combine responses in order to obtain more precise composite variates. Moreover, changing some experimental parameter could increase the slope of some responses and, hence, the precision of the potency estimates, as has been noticed for the observation time of chicken embryos. Finally, the system evaluation presented in this paper strictly concerns potency estimation as used in toxicity tests. No extrapolation to potential applications in other fields, like teratogenicity tests, seems legitimate.

Acknowledgments. This work was supported by grants 4.785.084.17 and 4.015.086.17 from the Swiss National Science Foundation.

1 Finney, D. J., Statistical Method in Biological Assay, 3rd edn. Charles Griffin, London 1978.
2 Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. A., Robust Statistic: the Infinitesimal Approach. Wiley, New York 1986.

3 Honegger, P., Biochemical differentiation in serum-free aggregating brain cell cultures, in: Cell Culture in the Neuroscience, pp. 223–243. Eds J. E. Bottenstein et al. Plenum, New York 1985.
4 Honegger, P., and Werffeli, P., Use of aggregating cell culture for toxicological studies. Experientia *44* (1988) 817–823.
5 Huber, P. J., Robust Statistics. Wiley, New York 1980.
6 Kučera, P., and Burnand, M. B., Routine teratogenicity test that uses chick embryos in vitro Terat. Carcin. Mutag. *7* (1987) 427–447.
7 Kučera, P., and Burnand, M.-B., Teratogenicity screening in standardized culture of chick embryo: effects of dexamethasone and diphenylhydantoin. Experientia *44* (1988) 827–833.
8 Marazzi, A., and Randriamiharisoa, A., ROBETH-ROBSYS: a software for robust statistical computing. Cahiers de recherche et documentation, Institut universitaire de médecine sociale et préventive, Bugnon 17, CH-1005 Lausanne (1985).
9 Marazzi, A., and Randriamiharisoa, A., ROBSYS: a package for robust statistical computing, *MEDINFO 86*, pp. 1111–1113. Eds R. Salamon and M. Jørgensen. Elsevier/North Holland, Amsterdam 1986.
10 Marazzi, A., and Randriamiharisoa, A., Statistical analysis of the parallel line and slope ratio assay using ROBSYS. Cahiers de recherche et documentation, Institut universitaire de médecine sociale et préventive, Bugnon 17, CH-1005 Lausanne (1987).
11 Schmid, B. P., and Cicurel, L., Application of the postimplantation rat embryo culture system to in vitro teratogenicity testing. Fd Chem. Toxic. *24* (1986) 623–626.
12 Schmid, B. P., and Cicurel, L., Postimplantation embryo culture for the assessment of the teratogenic potential and potency of compounds. Experientia *44* (1988) 833–840.
13 Sen, P. K., Robust statistical procedures in problems of linear regression with special reference to quantitative bioassay, I. Rev. int. statist. Inst. *39* (1971) 21–38.
14 Sen, P. K., Robust statistical procedures in problems of linear regression with special reference to quantitative bioassay, II. Rev. int. statist. Inst. *40* (1972) 161–172.

0014-4754/88/100857-17$1.50 + 0.20/0

Cryopreservation of parasites

J. Eckert

Institute of Parasitology, University of Zürich, Winterthurerstr. 266a, CH-8057 Zürich (Switzerland)

Summary. In this review, advances in cryopreservation of helminth parasites are reported. Our own studies demonstrate that metacestodes of *Echinococcus multilocularis* can be maintained in a viable state for at least 1–2 years by appropriate deep-freezing and storage in liquid nitrogen. Infective larvae of the nematode *Toxocara canis* cryopreserved for 1 week in liquid nitrogen were maintained after thawing in vitro in a chemically defined medium for 35 weeks. Although motility of previously deep-frozen larvae was reduced they produced secretory/excretory antigens of similar immunodiagnostic quality as those from unfrozen larvae. Whereas infective larvae of several species of trichostrongylids can be easily cryopreserved, the infective larvae of the cattle lungworm, *Dictyocaulus viviparus*, and muscle larvae of *Trichinella spiralis* are more sensitive to damage by subzero temperatures. Therefore, survival rates after cryopreservation are low, but improvement of the cooling schedules appears to be feasible. It is concluded that cryopreservation of certain stages of helminth and protozoan parasites is a useful technique for long-term storage of defined isolates, which can contribute considerably to reducing the number of experimental animals usually required for serial passages.

Key words. Cryopreservation; *Echinococcus multilocularis;* trichostrongylids; *Toxocara canis; Dictyocaulus viviparus; Trichinella spiralis;* alternatives to animal experimentation.

Introduction

Deep-freezing and storage in liquid nitrogen is a well-established and widely used technique for preserving protozoan parasites in a viable state[12, 16, 17, 22, 27–30]. In contrast, the application of this method to the metazoan helminths is still limited[12], although several species and stages have been cryopreserved successfully, for example the infective larvae of several species of trichostrongylids and strongylids[2, 4, 31, 34, 36], the microfilariae and/or infective larvae of *Brugia, Dirofilaria, Dipetalonema, Litomosoides, Wuchereria* and *Onchocerca*[9, 10, 11, 25, 26] and the sporocysts and schistosomula of *Schistosoma mansoni*[3, 13, 14, 18, 19, 23, 24, 35].

Reliable cryopreservation techniques for parasite stages can contribute considerably to the reduction of the number of experimental animals needed for strain maintenance in laboratories. In addition, these techniques would allow the preservation of viable parasite isolates and strains with interesting characteristics (drug resistance, high or low antigenicity etc.), and the establishment of parasite-banks for living reference material, and they would also facilitate storage for long periods and transport between laboratories[5, 6, 16]. Further, they could be applied in the production of live vaccines as shown by James and Dobinson[17] and Lewis et al.[23, 24] with schistosomula of *S. mansoni.*

In the following report we summarize some results of our recent research activities on cryopreservation of helminth parasites, which have been published in extenso elsewhere[1, 7, 32, 33].

Experiments and results

1. Cryopreservation of Echinococcus multilocularis metacestodes

The larval (metacestode) stage of *Echinococcus multilocularis*, which causes alveolar echinococcosis in humans, is